See Also:

The UBC School of Library, Archival and Information Studies Student Journal

2015 - Spring

Seeing Metadata: Representing Digital Collection Metadata

Evan Peter Williamson - evanpeterw@gmail.com

Keywords: digital collections, metadata, visual analytics, data visualization

Abstract:

Creators of digital content are often narrowly focused on each individual collection or project. In the face of the ever increasing scale of digital collections, it is challenging to step back and evaluate the metadata of the repository as a whole. This project outlines methods to visualize, evaluate, and use metadata in ways that overcome limitations inherent in the traditional management tools and workflows. The results demonstrate that visual analytic approaches have the potential to open up metadata from a new, more holistic perspective—freeing us from the item-level tabular view to grasp the larger patterns and structures in the data. Creating visual representations of metadata will be a valuable tool for management, analysis, and exploration.

Copyright:

Seeing Metadata: Representing Digital Collection Metadata

Evan Peter Williamson - evanpeterw@gmail.com

Evaluating the consistency of metadata describing digital collections is an essential task as the sheer number and diversity of items hosted online by cultural institutions increases. Digital Collections at the University of British Columbia (UBC) have grown to over 100,000 items of diverse genre, type, and format (UBC Library, n.d.). Divided into more than 50 collections, in-house metadata practices and standards have evolved over time. A close examination of item-level metadata reveals many inconsistencies and anomalies within and across collections. As Dushay and Hillmann (2003) observe, digital repositories face a challenge where:

> Instead of trained library workers applying well documented content standards to describe a relatively small number of resource formats, there are untrained people working largely in isolation (and without adequate documentation) to describe an increasingly complex array of resources. (2003, p.1)

Welcome to the metadata "Wild West"!

At UBC, the temporary student work force, evolving standards, and limitations in content management systems make quality control of metadata a cumbersome and difficult project. However, as aggregations of digital resources become larger, the impact of semantic and structural inconsistency in the metadata increases for users and administrators (Shreeves et al., 2005). Metadata that may have been suitable for a specific project is thrust into a much larger ecosystem, skewing findability, losing context, and becoming ambiguous. Improving interoperability and consistency of descriptive metadata was once primarily the concern of large scale aggregators, however, the effort has now shifted to individual institutions (Fenlon, Efron, & Organisciak, 2012). For administrators, regular audits of collection metadata can provide valuable feedback to improve current description practices while identifying and locating problems (Westbrook, Johnson, Carter, & Lockwood, 2012). Audits can also provide a clear analysis of the collection's coverage, helping to guide curation and ensuring that institutional mandates are met (Bahnemann, 2012). For the end-user, more consistent metadata will mean better access and discovery.

However, the growing scale of digital resources makes the process of managing metadata as a coherent whole increasingly challenging. One approach to solving the issue of the sheer volume is the application of visual analytic tools which efficiently combine the strengths of the human visual systems with the power of computing. As Ware describes, a well designed visualization enables the user to quickly comprehend a vast amount of information while perceiving patterns, detecting anomalies, and understanding the relation of large- and small-scale features in the data (Ware, 2013, p.3-4). Using a case study of metadata harvested from UBC's Digital Collections, this paper explores visual analytic approaches for auditing repository metadata.

Some previous research has outlined using visual analytics for metadata appraisal, driven mainly by aggregators seeking to convert data into more consistent and compliant forms. Dushay and Hillmann (2003) described using Spotfire DecisionSite, a visual analytics software, to efficiently evaluate large batches of XML metadata. A simple scatter plot with element names as vertical axis, records as horizontal axis, and XML encoding schema as colour/size allowed the researchers to quickly identify missing or incomplete values and improper encoding. Using similar visualizations, Nichols et al. (2008) designed a simple web-based tool that also produced statistics about the frequency of unique values. The tool was further developed as part of the Greenstone digital library software; however, the demonstration is no longer available at the URLs referenced in the article (as of December 2013). Nichols et al. (2009) reports on experiences using two online metadata analysis tools by institutional repositories in New Zealand. Fenlon et al. (2012) propose a series of statistical measures and visualizations as “administrator-oriented” tools to improve management of large aggregations of metadata.

To further explore the possibilities of metadata visualization to overcome limitations inherent in traditional management interfaces and workflows, this paper outlines five steps completed with the UBC Digital Collections metadata case study:

1. Harvest: remotely harvesting metadata from the repository.
2. Survey: using survey plots and other representations to visually evaluate usage of the metadata elements.
3. Refine: using clustering to refine the metadata for quality control and consistency.
4. Network: representing the metadata in network graphs to explore the usage of subject terms and the structure of relationships between collections.
5. Dashboard: using metadata to represent and navigate the collection as a whole.

Freed from the item-level tabular view, these representations offer methods for auditing consistency, locating problems, and better characterizing the scope of the repository. As the metadata is amended and refined, the visualizations become more meaningful representations of content. These clearer, more comprehensible pictures of the repository as a whole offer insight for both administrators and end-users, revealing patterns over huge numbers of items that would be impossible to perceive with conventional textual representations.

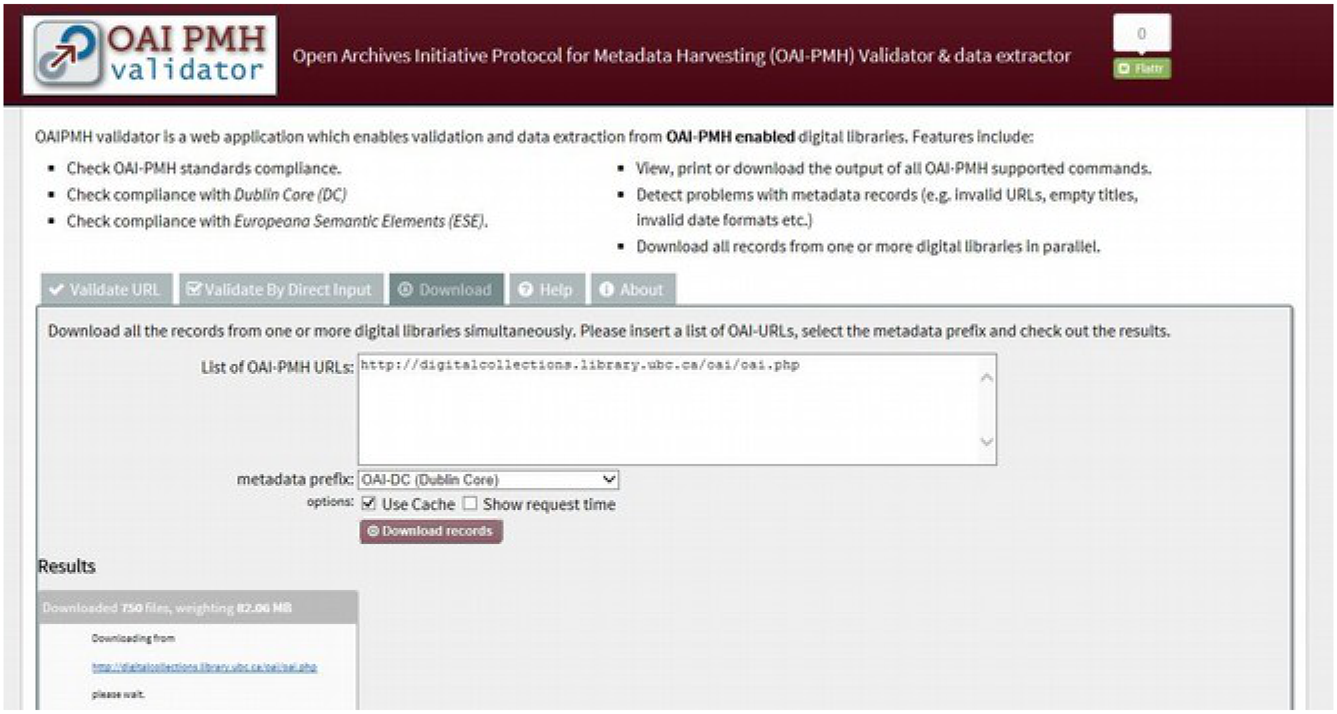
1. Harvest

The first step is, of course, to get data. UBC Digital Collections are built using CONTENTdm which allows two methods of exporting metadata:

1. Individual collection metadata can be exported from the web-based CONTENTdm Administration interface as a tab delimited TXT spreadsheet or as an XML file. Each item has the fields of the collection template plus a variety of machine-generated values. There is no utility to batch export more than one collection from the admin interface.
2. Full repository metadata can be harvested remotely via Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). OAI-PMH allows applications to request collection metadata from the server using a set of simple commands (Open Archives Initiative, 2002). Unless otherwise configured, the fields returned by the harvest are the 15 unqualified Dublin Core (DC) elements. CONTENTdm allows custom fields in each collection to be mapped to DC elements for the purpose of harvesting.

The first option can only be done in-house by employees with the proper permissions. Since it returns the customized metadata fields for individual collections and cannot be done en mass, it is less useful for assessing the repository as a whole. Remote harvesting, on the other hand, provides an opportunity to audit the “outside” view, to better assess interoperability. Since the harvest will return only the 15 unqualified DC elements, it provides a convenient and meaningful standard with the full repository as a frame of reference. The quality of the data set depends on the success of each collection's DC mapping, quickly revealing gaps that are not apparent in the smaller context of the custom templates.

There are a number of tools that make use of OAI-PMH which tend to be server-based and may require some scripting. However, a web-based tool called "OAI-PMH validator" makes a one time harvest easy (Banos, n.d). It is necessary to find and validate the OAI-PMH URL for the repository to ensure the server is working correctly. With “OAI-PMH validator”, simply enter the URL into the download box and the website begins harvesting the metadata (see Figure 1).


OAI PMH
validator
Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) Validator & data extractor
0
Flattr
✔ Validate URL
☑ Validate By Direct Input
Download
Help
About
Download records
Results

OAIPMH validator is a web application which enables validation and data extraction from **OAI-PMH enabled** digital libraries. Features include:

- Check OAI-PMH standards compliance.
- Check compliance with *Dublin Core (DC)*
- Check compliance with *Europeana Semantic Elements (ESE)*.
- View, print or download the output of all OAI-PMH supported commands.
- Detect problems with metadata records (e.g. invalid URLs, empty titles, invalid date formats etc.)
- Download all records from one or more digital libraries in parallel.

Download all the records from one or more digital libraries simultaneously. Please insert a list of OAI-URLs, select the metadata prefix and check out the results.

List of OAI-PMH URLs: http://digitalcollections.library.ubc.ca/oai/oai.php

metadata prefix: OAI-DC (Dublin Core)

options: ☑ Use Cache □ Show request time

Downloaded **750** files, weighting **82.06 MB**

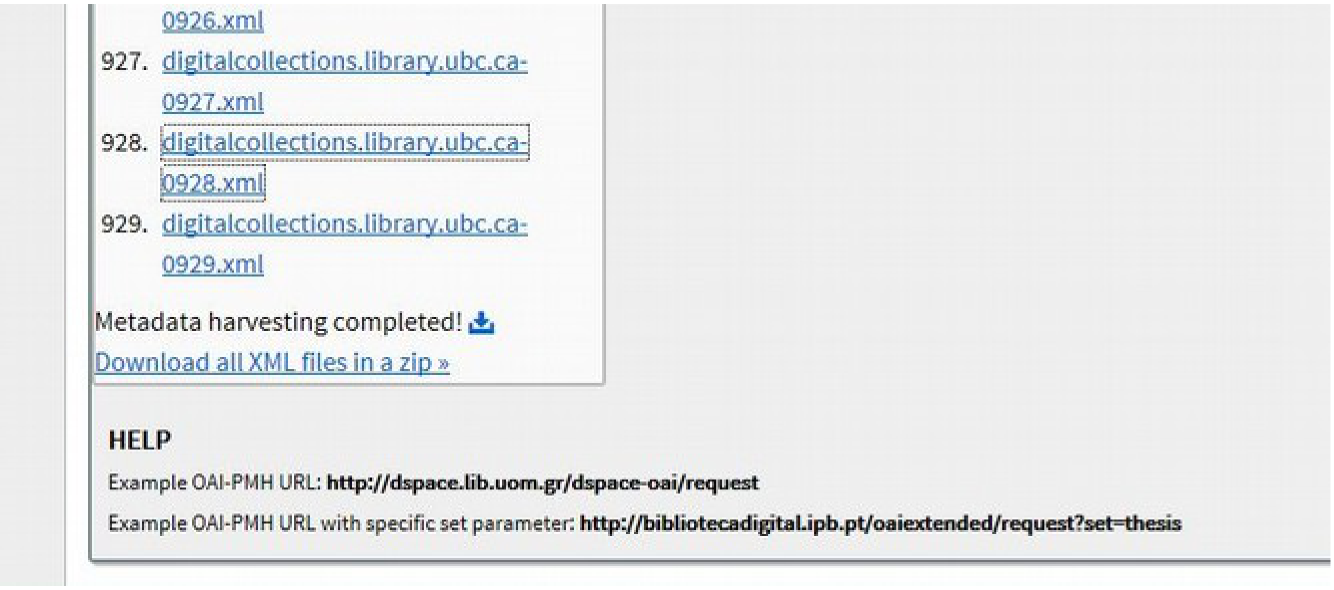
Downloading from

http://digitalcollections.library.ubc.ca/oai/oai.php

please wait.

Figure 1. Harvesting with OAI-PMH Validator.

OAI-PMH automatically streams the data as a series of XML files broken into manageable chunks. This can take awhile, as you may not know the full extent of the collections.



0926.xml

927. digitalcollections.library.ubc.ca-0927.xml
928. digitalcollections.library.ubc.ca-0928.xml
929. digitalcollections.library.ubc.ca-0929.xml

Metadata harvesting completed!

Download all XML files in a zip »

HELP

Example OAI-PMH URL: **http://dspace.lib.uom.gr/dspace-oai/request**

Example OAI-PMH URL with specific set parameter: **http://bibliotecadigital.ipb.pt/oaiextended/request?set=thesis**

Figure 2. OAI-PMH harvest complete.

An individual item in the resulting XML files looks like this:

```
<record><header><identifier>oai:digitalcollections.library.ubc.ca:ams/106</identifier
><datestamp>2013-04-09</datestamp><setSpec>ams</setSpec></header>
<metadata>
<oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"f
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
<dc:title><![CDATA[Blake Frederick, President]]></dc:title>
<dc:identifier><![CDATA[6529]]></dc:identifier>
<dc:identifier><![CDATA[2009.010.219]]></dc:identifier>
<dc:format><![CDATA[1]]></dc:format>
<dc:format><![CDATA[image/jpeg]]></dc:format>
<dc:type><![CDATA[Still Image]]></dc:type>
<dc:type><![CDATA[Photographs]]></dc:type>
```

```
<dc:creator><![CDATA[Deo, Gerald]]></dc:creator>
<dc:subject><![CDATA[AMS Executive]]></dc:subject>
<dc:date><![CDATA[2009]]></dc:date>
<dc:date><![CDATA[2010]]></dc:date>
<dc:format><![CDATA[3284x2448]]></dc:format>
<dc:description><![CDATA[Good]]></dc:description>
<dc:format><![CDATA[Yes]]></dc:format>
<dc:identifier><![CDATA[http://digitalcollections.library.ubc.ca/cdm/ref/collection/ams/id/106]]></dc:identifier></oai_dc:dc>
</metadata>
</record>
```

A harvest of UBC Digital Collections in November 2013 resulted in a batch of over nine hundred XML files with arbitrary breaks. It is easiest to convert these into tabular format using OpenRefine (http://openrefine.org). Simply select them all as a batch, highlight the record elements to parse, and generate a new project. The case study harvest resulted in a set of 73,000 records with the 15 DC elements.

A quick glance at this metadata reveals strange values, immediately revealing some issues with the DC mapping. When setting up a collection in CONTENTdm, custom fields can be mapped to qualified DC. Behind the scenes, this qualified DC is mapped to the unqualified DC returned by the harvest. Thus, some collections have multiple metadata fields (or none) mapped to a single DC element, creating multi-valued fields and empty fields inconsistently across the full Digital Collections. For example, the harvested “dc_identifier” field usually contains a mix of Catalog Call Number, Digital Identifier, and the unique CONTENTdm item URL (included since other identifiers are not necessarily unique). The “dc_date” field often contains several dates relating to both the original item and the digital object with no qualifiers.

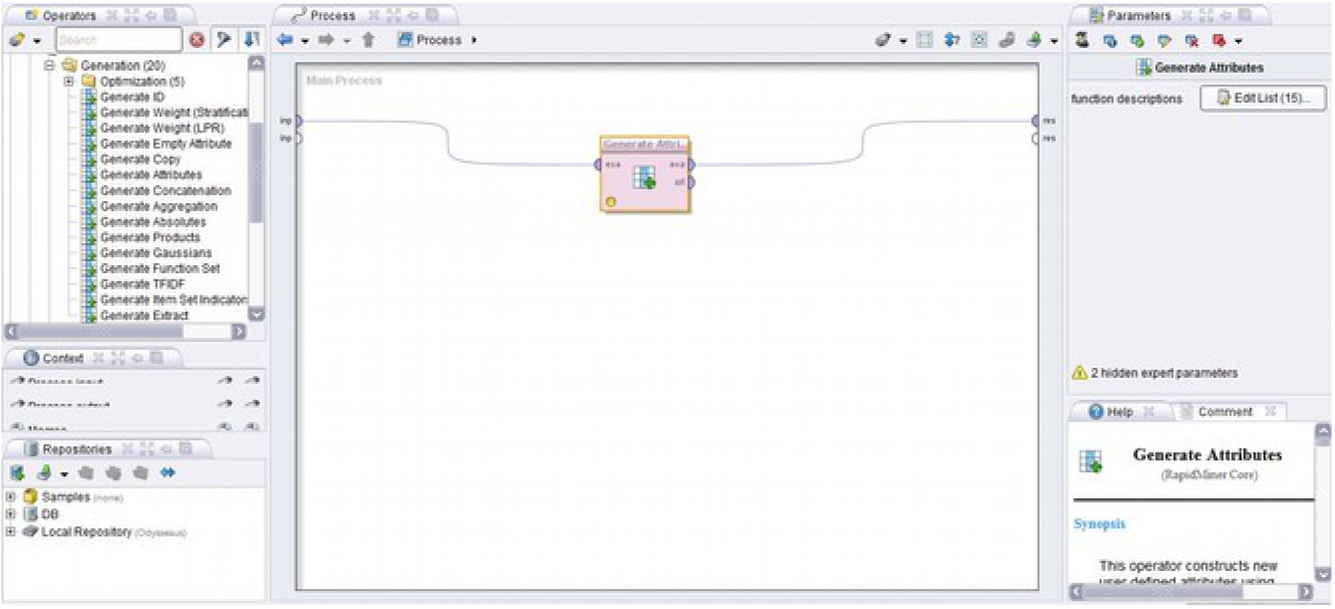
To finish preparing the data, each column was renamed to avoid spaces (e.g. "oai_dc:dc - dc:format" becomes "dc_format") and multi-valued cells were joined (each value was separated with a semicolon). This data set was exported from OpenRefine as a tab delimited file, then opened with Microsoft Excel and saved as a XLSX. This extra step is necessary because OpenRefine can not export XLSX, only XLS files which are limited to 65,000 rows.

2. Survey

As mentioned above, previous studies described a simple scatter plot enabling quick evaluation of the huge metadata sets harvested by aggregators (Dushay & Hillmann 2003; Nichols et al., 2008). These representations highlighted XML schema with multiple visual elements. This information is irrelevant in the case of UBC Digital Collections, since all harvested records follow the same DC schema. Instead, I wanted to create a survey plot of the character length of each field value. This visualization will give a greater sense of the composition and distribution of values in the fields, while still provide a sweeping overview of the data set.

To create the survey plot I used the popular data mining and analytics tool RapidMiner (http://rapidminer.com). In what should be taken as a tale of caution about the speed of change in a cutting edge (and potentially lucrative) field, RapidMiner underwent significant changes during the completion of this project prompted by an influx of investment money (Lunden, 2013). At the beginning of this project, the full RapidMiner application was offered for free use to individual non-commercial users by the German developer Rapid-I. In November 2013, the forums disappeared and the website was redirected. A new “Starter Edition” of RapidMiner6 was offered without charge, but has many limitations in place compared to the previous versions. Luckily, the “Community Edition” (the old RapidMiner5) became open source under the new “Business Source” model.

I selected RapidMiner because it was cost free, easy to use, and allows the creation of reusable processing routines and simple visualizations. RapidMiner's GUI has two "perspectives": *Design Perspective* is used to create flow charts that link together data input and processing steps; *Results Perspective* shows the final results of data processing and generates visualizations. Switching between views allows quick and efficient modification of processing routines while working out the kinks. This ability to quickly adapt queries and representations on the fly is a hallmark of visual analytics, efficiently augmenting human visual skills with data computation.



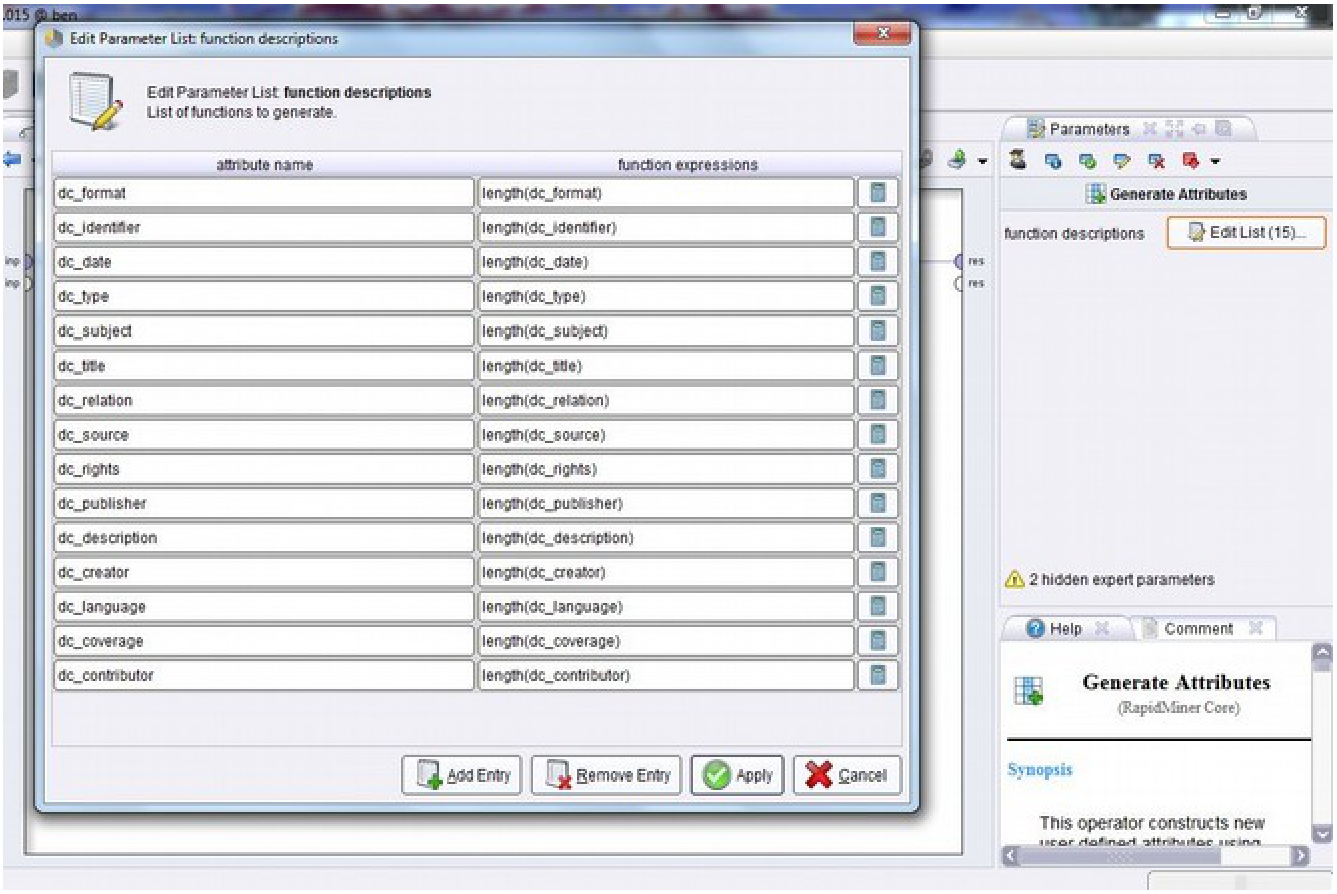
function descriptions: Edit List (15)...

2 hidden expert parameters

This operator constructs new

Figure 3. RapidMiner Design Perspective.

To create the survey plot, I set up a process in the design view (see Figure 3). On the left side of the screen is the "Operators" window—browse through the options and drag an operator into the "Process" window. Each process has an input and output node that can be connected to other operators. Different packages are available for download to add more operators (such as text processing or web mining). I set up a process to generate a new attribute equal to the number of characters in each value of the DC fields (see Figure 4). For example, a value of “An interesting item” in the dc_title field will be converted to 19. Since the fields are the same in every harvest, this process is reusable.



Edit Parameter List: function descriptions

Edit Parameter List: **function descriptions**
List of functions to generate.

attribute name	function expressions
dc_format	length(dc_format)
dc_identifier	length(dc_identifier)
dc_date	length(dc_date)
dc_type	length(dc_type)
dc_subject	length(dc_subject)
dc_title	length(dc_title)
dc_relation	length(dc_relation)
dc_source	length(dc_source)
dc_rights	length(dc_rights)
dc_publisher	length(dc_publisher)
dc_description	length(dc_description)
dc_creator	length(dc_creator)
dc_language	length(dc_language)
dc_coverage	length(dc_coverage)
dc_contributor	length(dc_contributor)

function descriptions: Edit List (15)...

2 hidden expert parameters

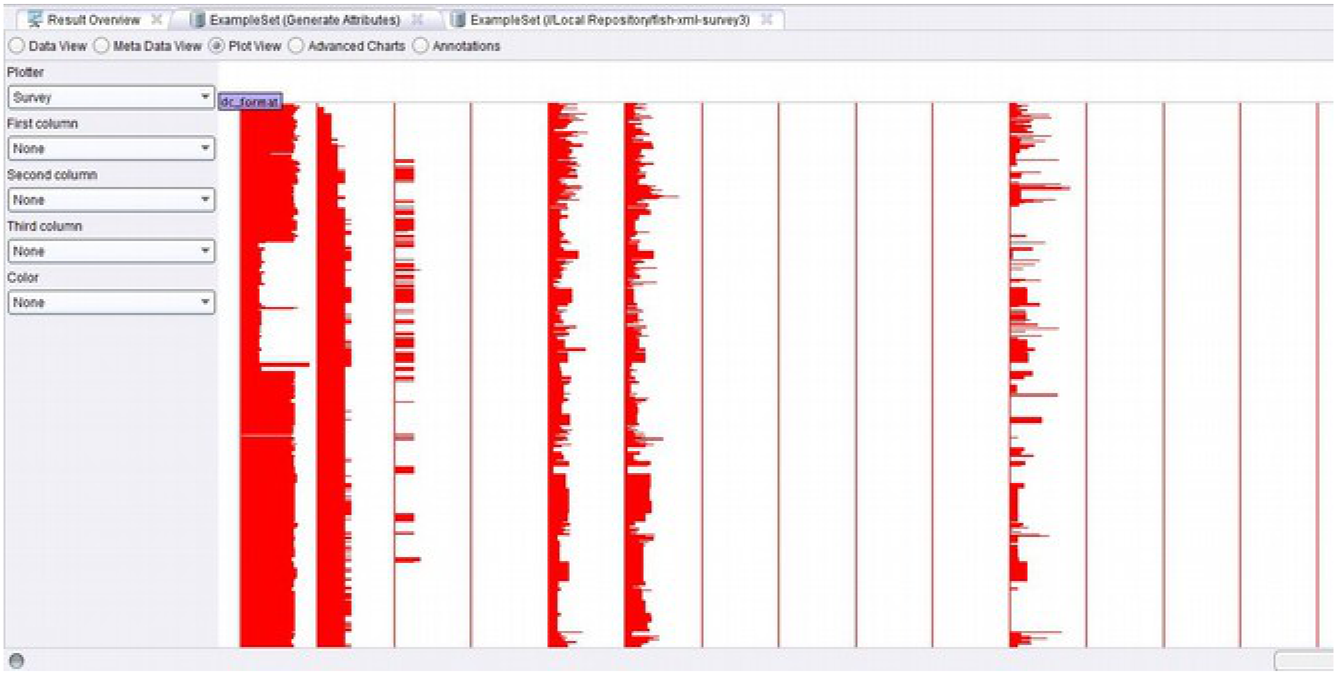
Generate Attributes
(RapidMiner Core)

Synopsis

This operator constructs new user defined attributes using

Figure 4. Setting parameters for the generate new attribute function.

Once the process is set up, import the metadata spreadsheet, click play, and the new data set arrives in the “results perspective” within seconds. Next, click on “plot view” and choose “survey plot” from the drop down menu. A survey plot of 73,000 items is created in practically no time at all!



Plotter
Survey
First column
None
Second column
None
Third column
None
Color
None

Figure 5. RapidMiner survey plot.

While the survey plot is simple and quick to produce, it has a number of frustrating limitations in RapidMiner5:

1. Slow scroll speed: While it generates nearly instantly, the full plot of 73k items is slow to interact with. Plots with less than 5k records could be scrolled through freely.
2. No Labels: There are no labels displayed on the columns or rows.
3. No details-on-demand: This plot does not link back to the underlying data. The documentation claims a details-on-demand feature is available, however, it was not present in my instance of the software.
4. Colour: This plot can be coloured using only an attribute that is included in the plot. Since the values are only the character length of the original records, this use of colour is meaningless. I would like to be able to colour

the plot using the original data values (such as Relation) to give further structure and analytic value to the representation.

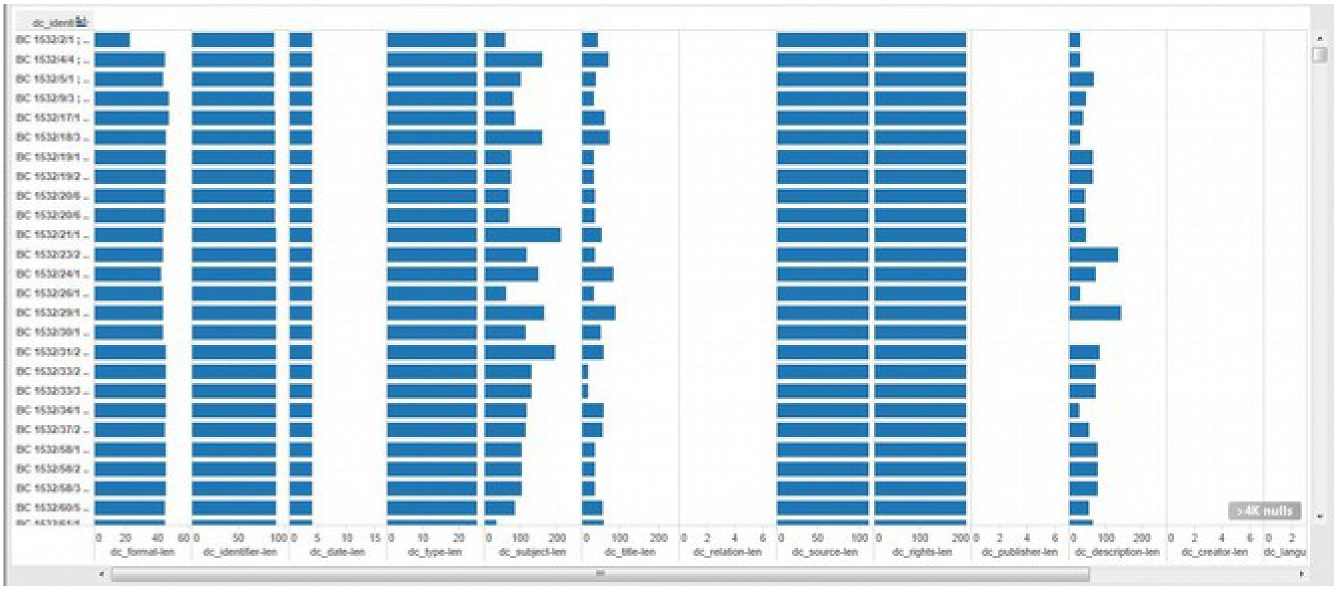
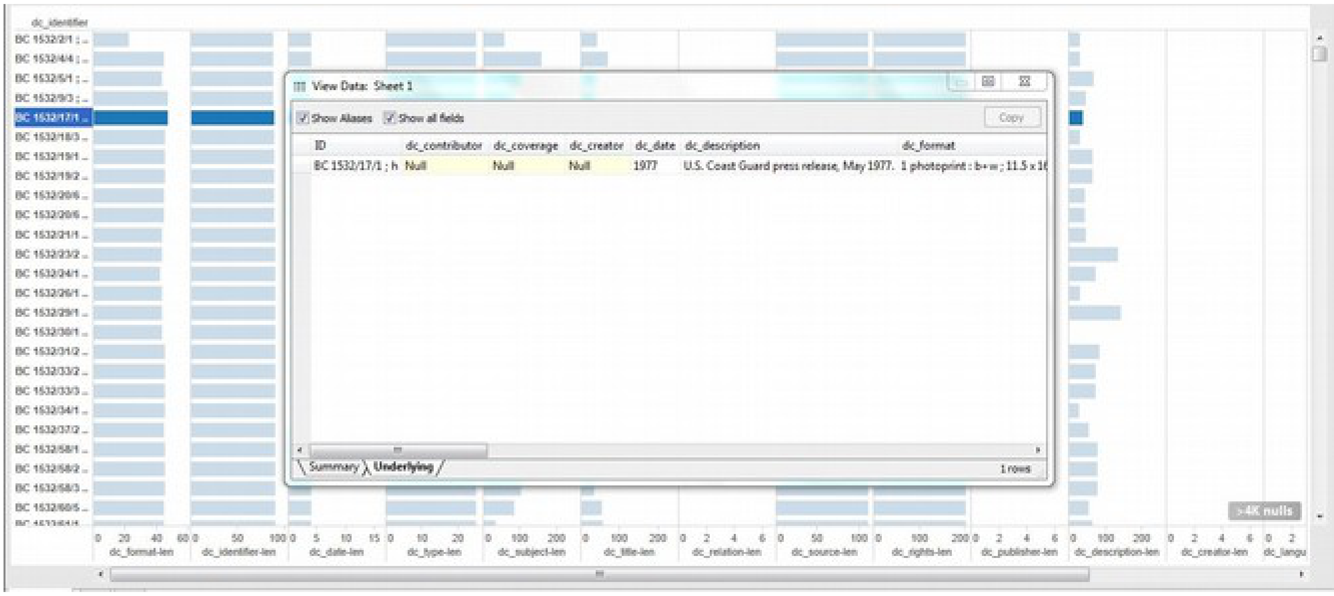


Figure 6. Survey plot in Tableau.

A similar plot can be created using Tableau (http://www.tableausoftware.com) that overcomes limitations 2, 3, and 4. However, setting up the visualization is considerably more work and is not reusable. Furthermore, the width of each item (even when reduced to the smallest size) is too large to allow for an efficient survey of the patterns in the collection at large and is very cumbersome to navigate.



ID	dc_contributor	dc_coverage	dc_creator	dc_date	dc_description	dc_format
BC 1532/17/1 ; h	Null	Null	Null	1977	U.S. Coast Guard press release, May 1977.	1 photoprint : b+w ; 11.5 x 1(

Figure 7. Tableau details-on-demand.

I evaluated a few other tools for creating this survey plot, but found none that could do it as quickly and easily as RapidMiner. In this case, I had exact specifications for a useful visualization in mind, but found it impossible to get all features from any one tool. In my search for other applications, it seemed that many open visual analytic developers are not creating complete standalone applications, but instead offer script libraries, tool kits, and frameworks to create custom tools with light programing skills. Since this survey plot process is highly specialized, a custom tool could better provide the desired functions while running more efficiently to handle the large data sets.

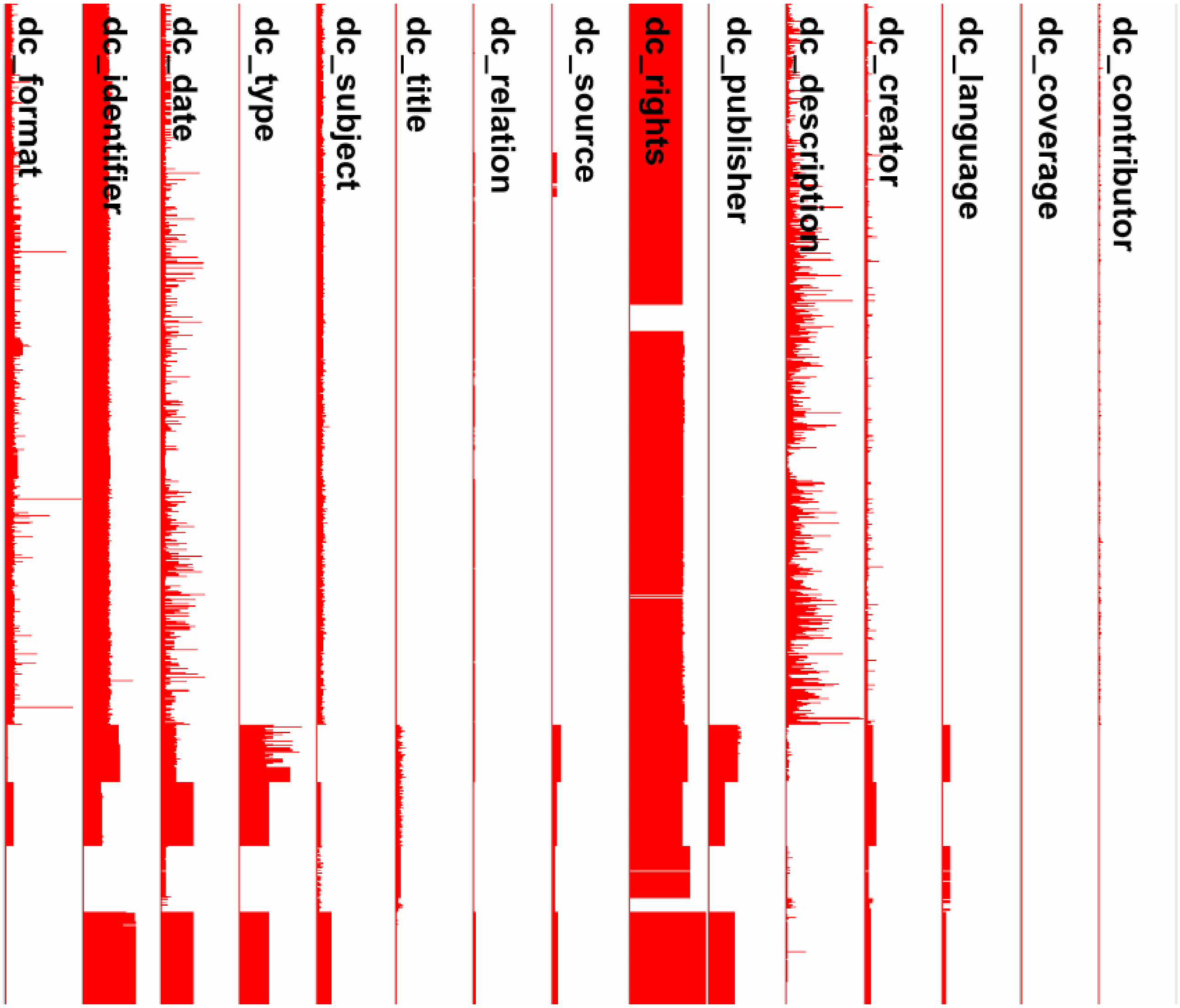
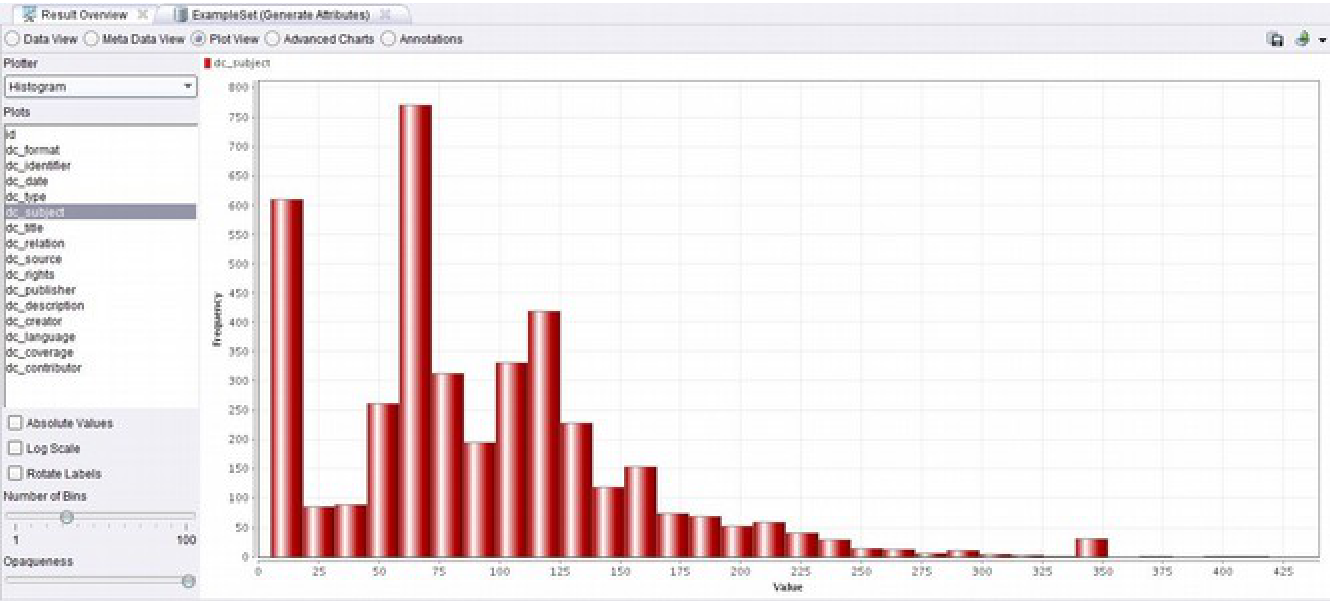


Figure 8. RapidMiner survey plot. A small section of the plot representing approximately 700 items with labels added for the DC elements.

For an example of how the survey plot can be used to understand metadata usage, examine Figure 8. This is a small section of the RapidMiner survey plot representing approximately 700 items. Each record is represented by a row one pixel tall, with one pixel representing each character in the DC elements. The fields appear in the order the harvest schema provides: Format, Identifier, Date, Type, Subject, Title, Relation, Source, Rights, Publisher, Description, Creator, Language, Coverage, and Contributor.

At first glance, it is immediately obvious that the Coverage field is not used at all—if a user is searching or browsing using geographic coverage, these items will be excluded. The Rights element is the most consistent, since the values are large blocks of text that are common for all items in one collection. However, there are several gaps that jumps out—we need to examine the underlying data to see which items are missing a rights statement! Almost every record has at least four pixels in the Date element, most likely a year in the original data. The larger values occur because multiple qualified DC date elements are mapped to the unqualified field. Several strange peaks appear in the Format and Type elements, which should be investigated since we would expect very consistent values in these fields. Over all, missing data, anomalies, and inconsistency are easily detected in the survey plot. The use or disuse of the elements is also immediately visible.

RapidMiner plots are extremely fast and simple, so after viewing the survey we can use some other visualizations to highlight different aspects of the data. For example, you might see a field that seems to have an unexpected amount of variation. Select “Histogram” from the plot menu, and select a field to see the distribution of lengths in the field overall. This visualization may reveal a pattern in the distribution that points to the source of the variation. It is most informative when applied to a single collection, since the variation in field length is easier to pin down (see Figure 9).

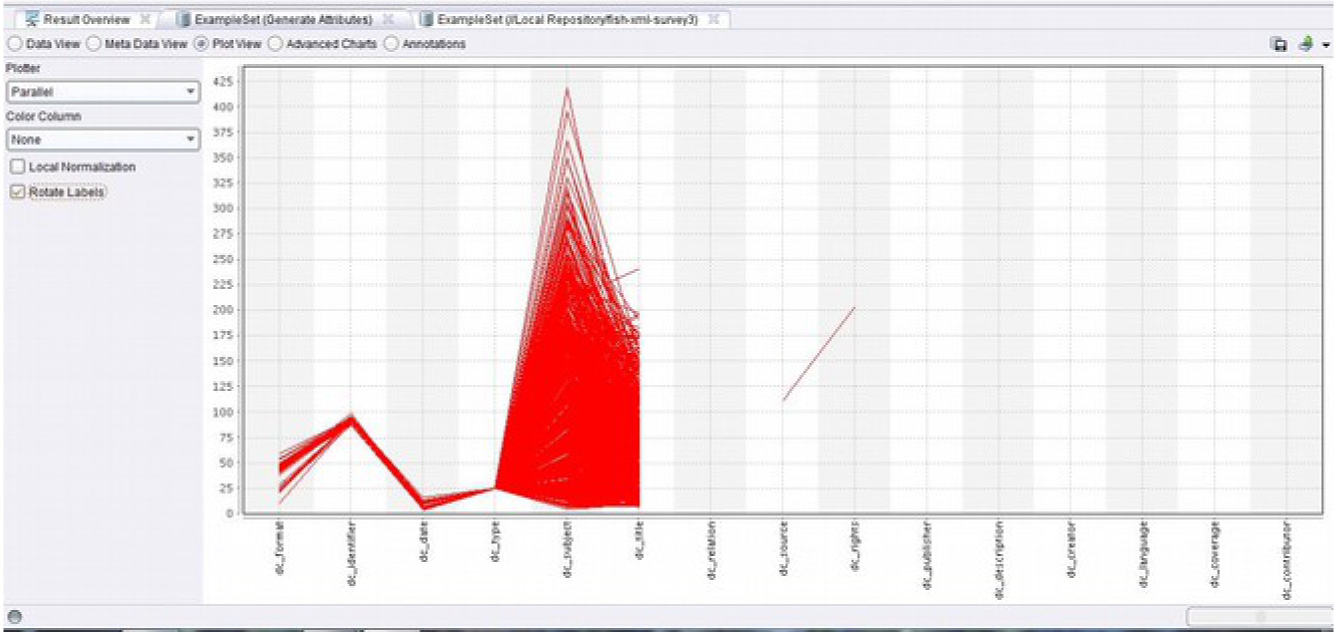

Plotter: Histogram

Plots: dc_subject

☐ Absolute Values
☐ Log Scale
☐ Rotate Labels

Figure 9. RapidMiner histogram of dc_subject.

Another option that has promise is a parallel plot. It gives a quick sense of the overall use of the fields in a very condensed visualization - a one page report card. In this application it is crowded, there is no actual tracking the individual items along the lines—just a quick impression of the entire data set at a glance. This plot takes more time to generate in RapidMiner and is quite basic. A more fully functional tool could create a larger scale, more beautiful version of this visualization. It would be helpful to colour the lines by the original source field to highlight logical groupings and see how coherent the individual collections are within the larger context.



Plotter: Parallel

Color Column: None

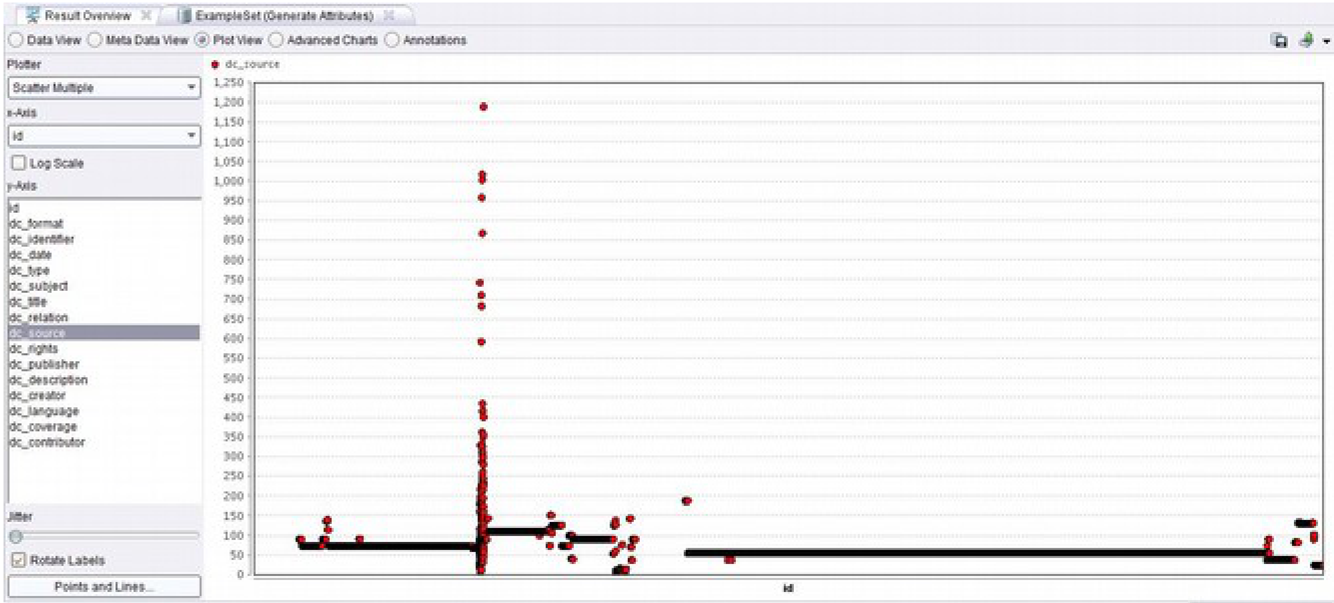
□ Local Normalization

☑ Rotate Labels

Figure 10. RapidMiner parallel plot of Fisherman Publishing metadata.

For example, looking at the parallel plot of Fisherman Publishing (see Figure 10), notice that the dc_format is of varying length (from about 10 to 60 characters). This is concerning because they are all photographs and should all have the same value. Type, source, and rights are all exactly the same, as they should be. Identifier and date are not all the same, but are tightly packed which makes sense for the underlying data. For example, the date field ranges from four characters for a year (e.g. 1950) to ten characters for a full date (e.g. 1950-05-04). Subject is highly varied, which is expected since some items are assigned multiple subject terms while others have none. The seven unused elements are immediately obvious.

Going further, a quick modification to the process can add an ID field to the lengths using the original identifiers. We can then use the ID field as an x-axis domain to plot against. RapidMiner's "ScatterMultiple" makes it easy to plot any or all of the fields. Trying to select all the fields is too crowded to be useable, but clicking through the fields one by one provides a quick overview of each. This has the advantage of isolating the use of each element, but the disadvantage of not seeing it as a complete picture as in the survey. However, the plot is instantaneous, so it is very easy to flip through.



Plotter
Scatter Multiple

x-Axis
id

□ Log Scale

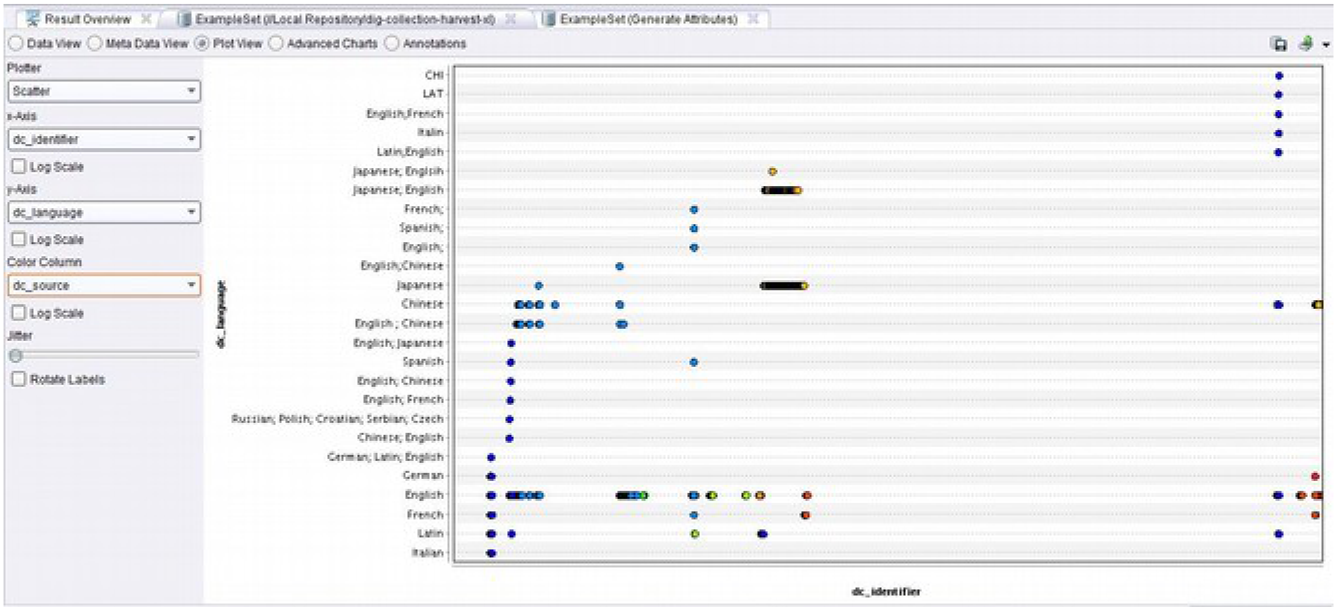
y-Axis
id
dc_format
dc_identifier
dc_date
dc_type
dc_subject
dc_title
dc_relation
dc_source
dc_rights
dc_publisher
dc_description
dc_creator
dc_language
dc_coverage
dc_contributor

☑ Rotate Labels

Figure 11. RapidMiner scatter plot of DC Source.

For example, in a scatter plot of dc_source for the full UBC Digital Collections (see Figure 11), we can quickly see that most collections have the field filled, and they are fairly consistent values. Source should generally correspond to the repository name (e.g. UBC Rare Books and Special Collections), so we expect consistent values. However, we see one giant spike and two empty spaces. A peek back at the data (details-on-demand would make this very functional and quick evaluation tool) shows that McCormick Maps collection was using the source field incorrectly with long URL references to outside resources, thus creating the spike. The empty spaces are the AMS Photograph Collection and the Tairiku Nippo Newspaper Collection which did not have any field mapped to Source.

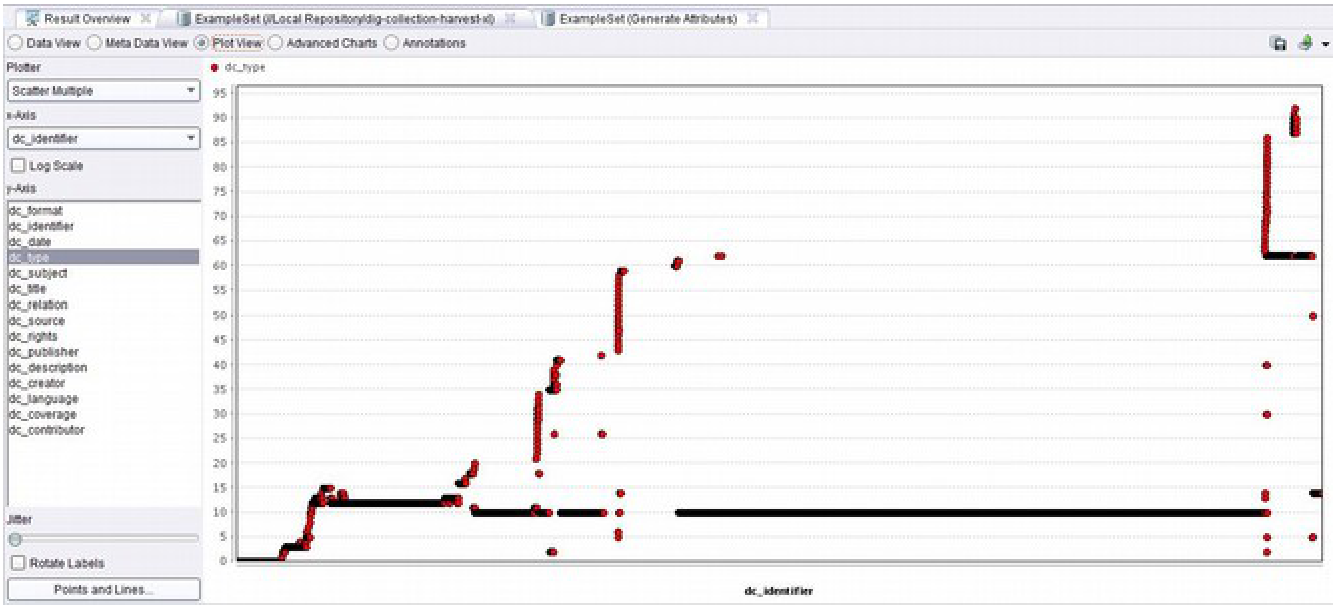
It is worth noting that RapidMiner is flexible enough to plot the data in its raw state as well. For example, a plot of dc_language shows the distribution of language values (see Figure 12). The element is often null, thus the plot is sparse, and the values reveal issues with consistency. If refined, the representation would provide insight into the languages available in the repository.



Plotter
Scatter
x-Axis
dc_identifier
☐ Log Scale
y-Axis
dc_language
☐ Log Scale
Color Column
dc_source
☐ Log Scale
Jitter
☐ Rotate Labels

Figure 12. RapidMiner scatter plot of raw values in the DC Language element.

A plot of dc_type gives us an idea of the distribution of item genre/type—the longest horizontal line represents "photographs ; still image" (47,071 items), with the next longest being "Still Image" (8,112 items), "still image ; photographs" (2,958 items), and "Still Image" (2,792 items) (see Figure 13). These values point out some of the consistency problems!


Result Overview
ExampleSet (//Local Repository/dig-collection-harvest-x)
ExampleSet (Generate Attributes)
○ Data View ○ Meta Data View ● Plot View ○ Advanced Charts ○ Annotations
Jitter
□ Rotate Labels
Points and Lines...
● dc_type
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
dc_identifier

Plotter
Scatter Multiple
x-Axis
dc_identifier
□ Log Scale
y-Axis
dc_format
dc_identifier
dc_date
dc_type
dc_subject
dc_title
dc_relation
dc_source
dc_rights
dc_publisher
dc_description
dc_creator
dc_language
dc_coverage
dc_contributor

Figure 13. RapidMiner scatter plot of raw values in the DC Type element.

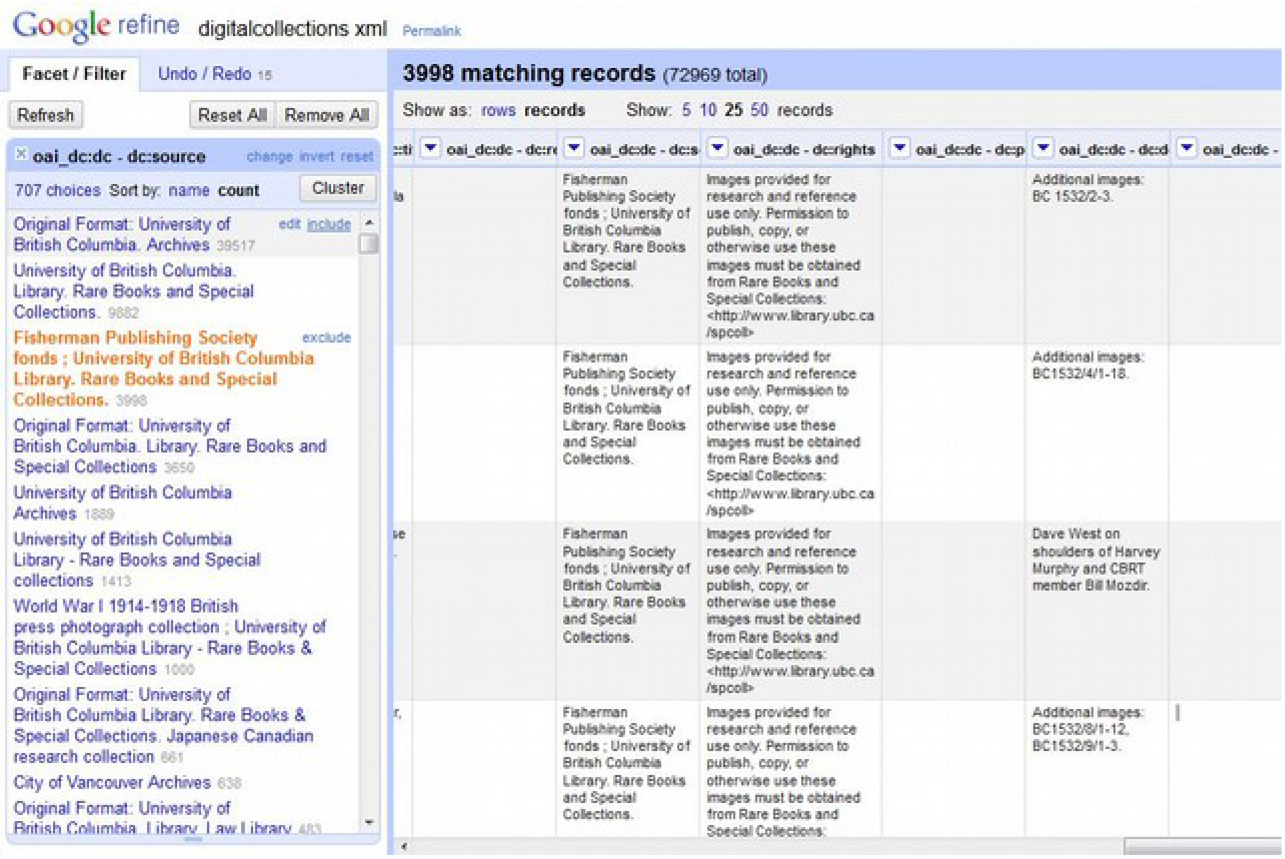
Since the axis are discrete non-numeric entities and the range is too large to distinguish the individual values, these representations give a high-level sense of structure. A thorough knowledge of the collections is necessary to interpret the patterns.

The ideal survey provides a quick, yet detailed evaluation of large data sets that are otherwise difficult to navigate and conceive of holistically. However, as implemented in the tools I used, it wasn't as efficient and informative as it should be. In this case, the RapidMiner ScatterMultiple plot seems the simplest method to quickly evaluate the harvested metadata, but each type of visualization offers different insights. These visualizations could be valuable to help improve metadata quality. They allow viewing the repository as a whole and reveal the problems introduced by mapping to DC. By highlighting significant anomalies or missing data, it helps reveal where quality control is most needed, allowing better prioritization of effort. A visualization with details-on-demand would also be an effective method of finding the needle in the haystack—it would allow us to quickly see which collections have issues, isolate the problem items, and fix them. Stepping back from the tabular values by using this abstracted visualization gives us an opportunity to see how the inconsistencies are impacting the collection as a whole.

3. Refine

As the example of dc_type demonstrates above, simple formatting errors that are not noticeable at the individual item level are a major challenge in the aggregated metadata. The four versions of “Photograph ; Still Image” all refer to the same concept, but are completely different values as data disrupting the collation of these related items. Most of the UBC Digital Collections metadata is based on existing library catalogue entries or is formatted following relevant cataloguing standards. However, these standards are not designed to be machine-readable, focusing on representing a single item rather than providing navigational facets as we are accustomed to using online. For example, the field Extent often has values such as “1 photograph : black & white, 8x10”, representing a series of related values constructed in a syntax that can not be sorted in any useful way. Similarly, square brackets are used to represent a title or date provided by the archivist or cataloguer—these qualifying symbols are expressive to information professionals (much less so to the public), but prevent proper sorting or representation of a complete series of values.

For the purposes of the case study, these inconsistencies and multivalued fields need to be cleaned up as much as possible to enable visualization of the collection as a whole. In a production context, repositories will want to refine metadata in a similar fashion to improve interoperability and consistency in a unified discovery environment. The best tool for this job is OpenRefine, a free, open source application for data wrangling that is powerful and easy to start using. OpenRefine is column centric: the interface displays a small sample of rows and navigation via facets or other operations run on a specific column (see Figure 14). The application is incredibly fast at large scale manipulations and navigation, but is not designed for individual data entry.



3998 matching records (72969 total)

Show as: rows records   Show: 5 10 25 50 records

oai_dc:dc - dc:source

707 choices Sort by: name count

Original Format: University of British Columbia. Archives	39517
University of British Columbia. Library. Rare Books and Special Collections.	9882
Fisherman Publishing Society fonds ; University of British Columbia Library. Rare Books and Special Collections.	3998
Original Format: University of British Columbia. Library. Rare Books and Special Collections	3650
University of British Columbia Archives	1889
University of British Columbia Library - Rare Books and Special collections	1413
World War I 1914-1918 British press photograph collection ; University of British Columbia Library - Rare Books & Special Collections	1000
Original Format: University of British Columbia Library. Rare Books & Special Collections. Japanese Canadian research collection	661
City of Vancouver Archives	636
Original Format: University of British Columbia. Library. Law Library	483

:ti	oai_dc:dc - dc:r	oai_dc:dc - dc:s	oai_dc:dc - dc:rights	oai_dc:dc - dc:p	oai_dc:dc - dc:d	oai_dc:dc -
la		Fisherman Publishing Society fonds ; University of British Columbia Library. Rare Books and Special Collections.	Images provided for research and reference use only. Permission to publish, copy, or otherwise use these images must be obtained from Rare Books and Special Collections: <http://www.library.ubc.ca/spcoll>		Additional images: BC 1532/2-3.	
		Fisherman Publishing Society fonds ; University of British Columbia Library. Rare Books and Special Collections.	Images provided for research and reference use only. Permission to publish, copy, or otherwise use these images must be obtained from Rare Books and Special Collections: <http://www.library.ubc.ca/spcoll>		Additional images: BC1532/4/1-18.	
se		Fisherman Publishing Society fonds ; University of British Columbia Library. Rare Books and Special Collections.	Images provided for research and reference use only. Permission to publish, copy, or otherwise use these images must be obtained from Rare Books and Special Collections: <http://www.library.ubc.ca/spcoll>		Dave West on shoulders of Harvey Murphy and CBRT member Bill Mozdir.	
r,		Fisherman Publishing Society fonds ; University of British Columbia Library. Rare Books and Special Collections.	Images provided for research and reference use only. Permission to publish, copy, or otherwise use these images must be obtained from Rare Books and Special Collections:		Additional images: BC1532/8/1-12, BC1532/9/1-3.	

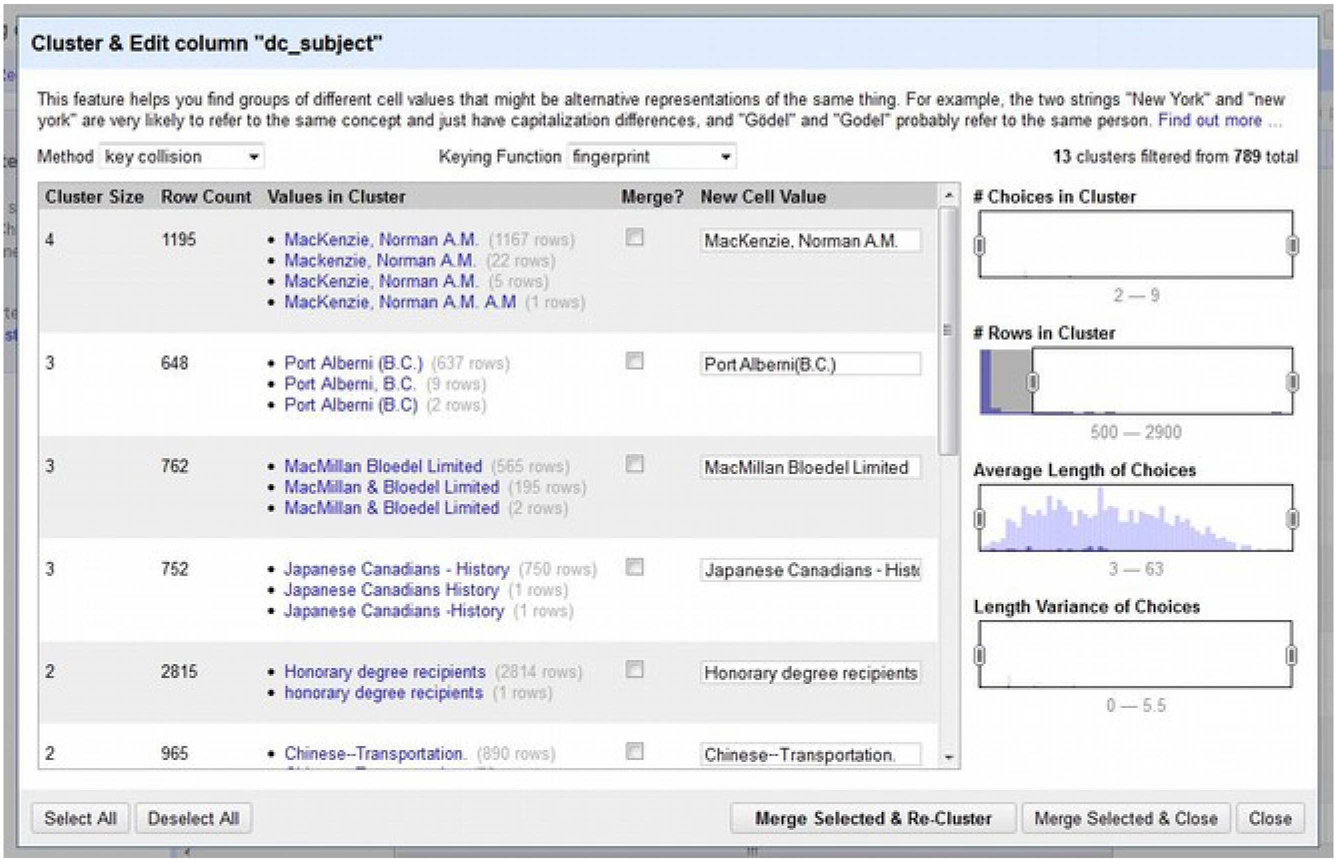
Figure 14. OpenRefine. Full harvested metadata in OpenRefine with a text facet on Source and Fisherman Publishing selected returns 3,998 records out of the 72k total.

OpenRefine can help automate refining the metadata by clustering values that are semantically the same, but are currently separated by formatting and spelling inconsistencies. Using the full harvested UBC Digital Collections metadata, I followed this procedure:

1. *Split multivalued cells* on all the columns using semicolon: Values that appear in various fields are often multivalued and use a semicolon to

separate the different subjects. Separating the values allows a text facet of the individual subjects, rather than the long strings of multiple values.

2. *Trim leading and trailing white space* and *to lowercase:* These operations quickly cut many formatting anomalies.
3. *Text facet: G*roups all the common values together as a "facet" which allows quick navigation through the data to assess its contents and distribution. Errors in the facets are often immediately obvious, editing the facet changes the entire batch of values.
4. *Text facet cluster:* Several clustering algorithms are available to offer suggestions about which text facets are actually the same, just formatted differently or spelled incorrectly. Using each algorithm surfaces different types of anomalies, although outside of “key collision fingerprint”, the rate of false positives gets increasingly higher (OpenRefine Wiki, n.d.). The tool offers the clusters and a new single value (see Figure 15). Since the metadata is handmade (often a controlled vocabulary was not well enforced) the clusters often reveal five or more variants in spelling and punctuation (e.g. black and white, Black and White, Black + White, Black & White, B+W, B&W, B & W, etc.). While these anomalies are easy to overlook and understand to a human at the item level, in analyzing data or navigating through large collections they cause significant issues.



Cluster & Edit column "dc_subject"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more ...

Method key collision Keying Function fingerprint 13 clusters filtered from 789 total

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
4	1195	• MacKenzie, Norman A.M. (1167 rows) • Mackenzie, Norman A.M. (22 rows) • MacKenzie, Norman A.M. (5 rows) • MacKenzie, Norman A.M. A.M (1 rows)	□	MacKenzie, Norman A.M.
3	648	• Port Alberni (B.C.) (637 rows) • Port Alberni, B.C. (9 rows) • Port Alberni (B.C) (2 rows)	□	Port Alberni(B.C.)
3	762	• MacMillan Bloedel Limited (565 rows) • MacMillan & Bloedel Limited (195 rows) • MacMillan & Bloedel Limited (2 rows)	□	MacMillan Bloedel Limited
3	752	• Japanese Canadians - History (750 rows) • Japanese Canadians History (1 rows) • Japanese Canadians -History (1 rows)	□	Japanese Canadians - Histo
2	2815	• Honorary degree recipients (2814 rows) • honorary degree recipients (1 rows)	□	Honorary degree recipients
2	965	• Chinese--Transportation. (890 rows)	□	Chinese--Transportation.

Figure 15. OpenRefine clustering on DC Subject.

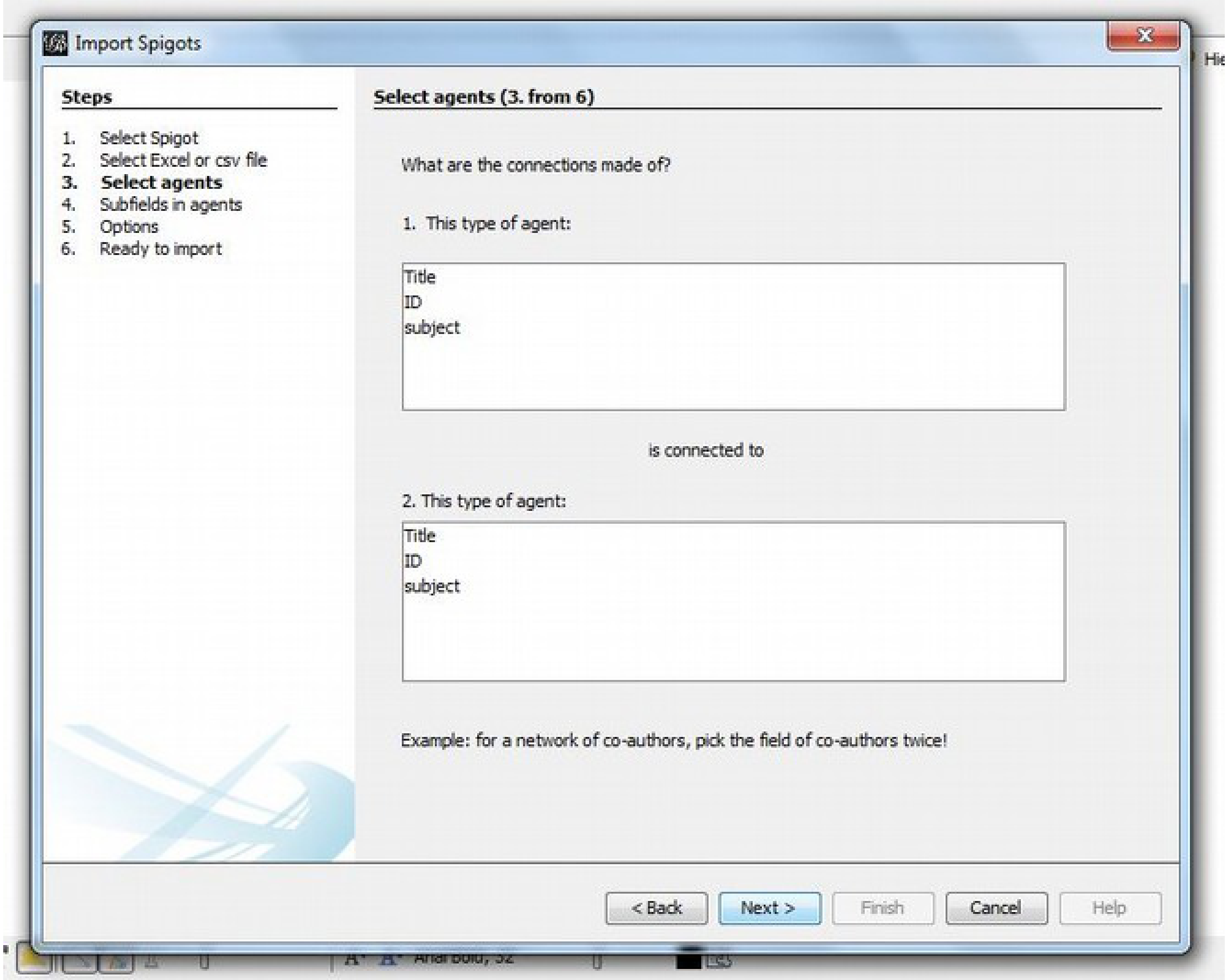
Looking at the metadata using OpenRefine provides a unique view that overcomes some of the limitations of the CONTENTdm client used to create the collections. The client is focused on individual item description and content creation rather than actual content management. It allows for working on sets of only 1000 items at a time, has poor sorting capabilities, and clumsy navigation. Only a small group of items can be viewed on the screen at one time. The interface limits the ability to visualize and correct metadata from a holistic view point. In contrast, OpenRefine enables users to quickly surface patterns and navigate across the entire collection. Both Fenlon et al. (2003) and Nichols et al. (2008) use tables giving summary statistics about the number of unique values in each field to help with their metadata evaluations. OpenRefine's text facet feature offers similar functionality, but provides more powerful options to navigate and ultimately fix the values.

See Also:
2015 - Spring

4. Network

Since fields such as Subject have overlapping values within and across collections, the metadata implies a variety of relationships between items within the repository. For example, Fisherman Publishing is a collection of 3,998 digitized photographs. 3,259 of the items have at least one subject term, with 2,006 having two or more. Furthermore, the Subject and Personal Names fields overlap considerably. If the fields are combined into a single column, only eight items do not have a subject. Thus, an item can be seen as related to another if they share one or more of the same subject terms—or a subject is related to another subject if they describe a common item. Since the number of items is large, the number of subjects is large, and items can have multiple subjects, it is impossible to grasp these complex relationships in tabular form. However, the links between these entities can be represented as a network graph to visualize and explore the structure.

Gephi (https://gephi.org) is an open source tool that builds beautiful and detail rich network graphs with many community designed plugins and extensions. Once you get the hang of the interface, it is a joy to navigate and interact with the graph. Calculating basic relationship data from a refined metadata set is simplified by the "Excel/csv converter to network" plugin for Gephi (Levallois, n.d.). This “Import Spigot” automatically generates the necessary node and edge tables from an existing spreadsheet file for a simple relationship specified by the user (see Figure 16).



Steps

1. Select Spigot
2. Select Excel or csv file
3. **Select agents**
4. Subfields in agents
5. Options
6. Ready to import

Select agents (3. from 6)

What are the connections made of?

1. This type of agent:

Title
ID
subject

is connected to

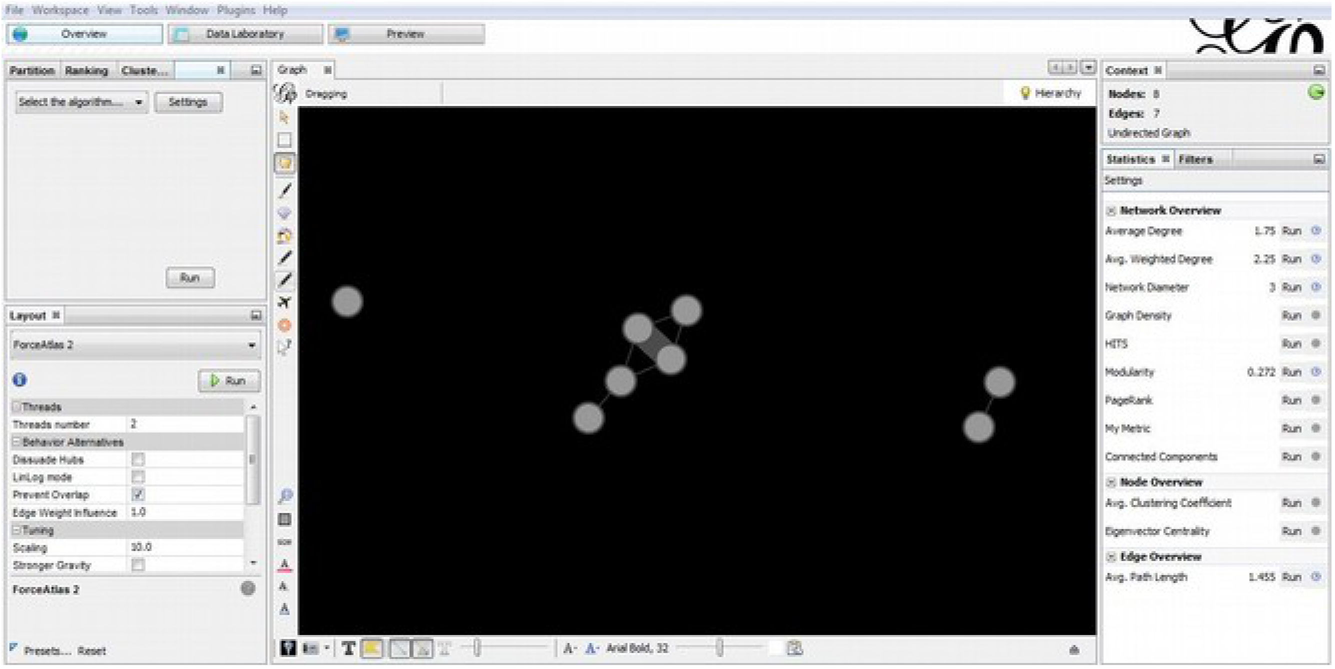
2. This type of agent:

Title
ID
subject

Example: for a network of co-authors, pick the field of co-authors twice!

Figure 16. Gephi import spigot. Using the "Excel/csv converter to network" plugin to create network data.

Starting with the refined UBC Digital Collections metadata, I used the Excel converter plugin to generate network data (node and edge tables) based on several possible relationships. For example, creating a connection from dc_subject to dc_subject results in 23,673 nodes (representing the unique subject terms) and 48,627 edges (representing relationships where two subject terms describe the same items). A connection of dc_subject to dc_relation results in 19,416 nodes (while there are 23,673 unique subject terms, many only describe items with a null relation field) and 67,538 edges. I am interested in visualizing the use of subject terms because these relationships reveal information about the application of metadata within collections as well as the character of the contents.



ForceAtlas 2

Threads	
Threads number	2
Behavior Alternatives	
Dissuade Hubs	□
LinLog mode	□
Prevent Overlap	☑
Edge Weight influence	1.0
Tuning	
Scaling	10.0
Stronger Gravity	□

Nodes: 8
Edges: 7
Undirected Graph

Network Overview	
Average Degree	1.75
Avg. Weighted Degree	2.25
Network Diameter	3
Graph Density	
HITS	
Modularity	0.272
PageRank	
My Metric	
Connected Components	
Node Overview	
Avg. Clustering Coefficient	
Eigenvector Centrality	
Edge Overview	
Avg. Path Length	1.455

Figure 17. Gephi interface. The central Graph pane displays a test data set using the ForceAtlas2 layout.

Working from these network data sets I followed a basic workflow in Gephi:

1. Run a force directed algorithm: The force directed algorithm breaks up a random layout of nodes following specific parameters to create clusters of related nodes. Basically, for each node the algorithm calculates a “force” attracting it towards related nodes and repulsing it from unrelated ones. Each node then moves a set amount based on the calculated force. This is repeated round after round until the layout reaches an equilibrium or is paused by the user (see Figure 17). Since the subject graphs are large, I used ForceAtlas2, OpenOrd, and YifanHu which can efficiently cluster very large networks using home PC resources.
2. Colour and rank the nodes: After completing one of the force directed layouts, other visual elements can be customized to represent different

features of the data. From the Partition window the nodes can be coloured using a variety of properties and stats. For example, dc_relation and dc_subject nodes can be represented by different colours. Using the Ranking window, the relative sizes of nodes can be adjusted based on their properties. For example, ranking nodes based on degree will highlight the most used subject terms.

3. Rank the labels: To ease exploration of the graph and quickly identify the central components, I rank the node labels by degree (in the Ranking window). Making the minimum size very small allows the labels to be left on while browsing the graph: they will only be visible when zooming in, thus it will not distract users from the high-level view.

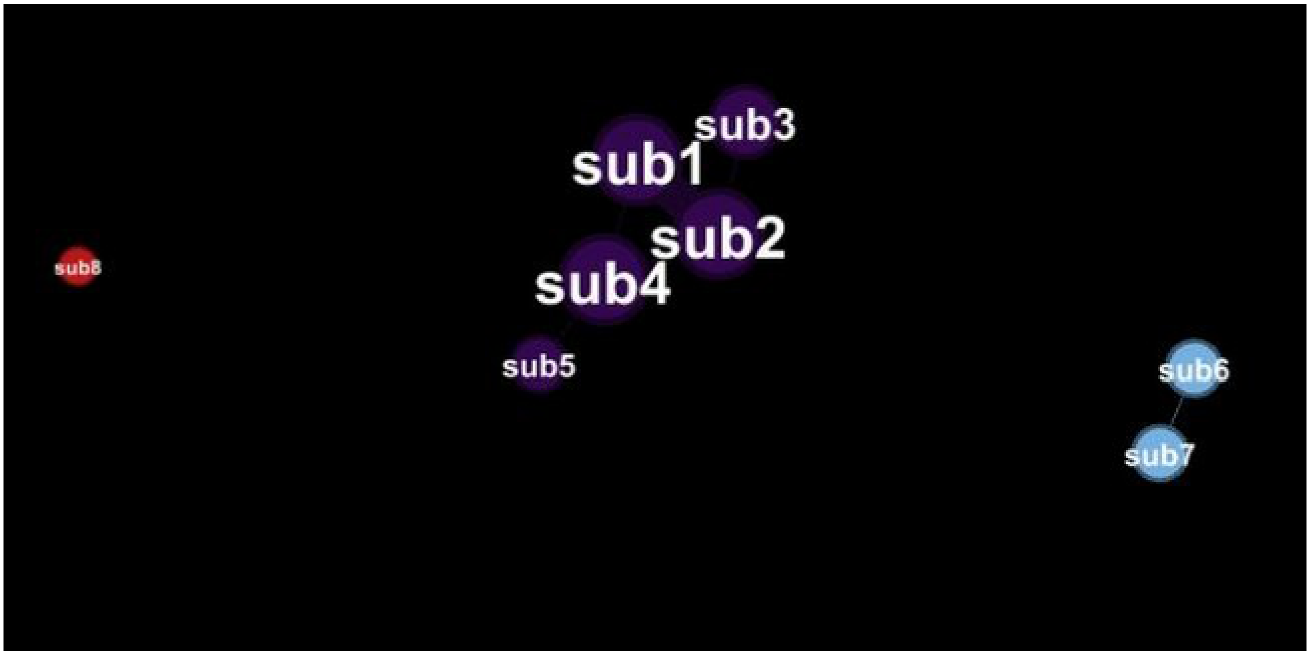


Figure 18. Gephi network graph. A test data set coloured by modularity class with ranked labels.

This creates a network graph that is easy to navigate and interact with. Nodes can be selected to highlight the underlying data or to start a new graph. The parameters of every visual element are adjustable (labels, edges, nodes, background, highlights, filters, etc.)—the options are endless and amazingly fluid.

The force directed layouts applied to the networks based on Subject reveal a central cluster, with many complex relationships, and a ring of outliers with few connections. A few distinct groups break away from the main clusters, with balanced links to each other and no relations to anything else—these are often a single item with many unique subjects that are not used else where. The various layouts are fascinating, but interpreting them can be challenging. It is necessary to understand both the relation of the original data to the nodes and edges tables, and the nature of the functions which pushed the layout.

For example, the graph in Figure 19 is based on links between DC Subject and Relation, clustered using the OpenOrd layout, and coloured by node type. Let us examine how it can be interpreted and what it reveals about metadata practice. The Relation field in UBC Digital Collections is typically used as the qualified DC field “relation—is part of” which describes the collection or fonds the item belongs to. The unique dc_relation values are represented by blue nodes in this network graph. The Subject field contains one or more subject terms used to describe the digital object. The unique dc_subject terms are represented by red nodes.

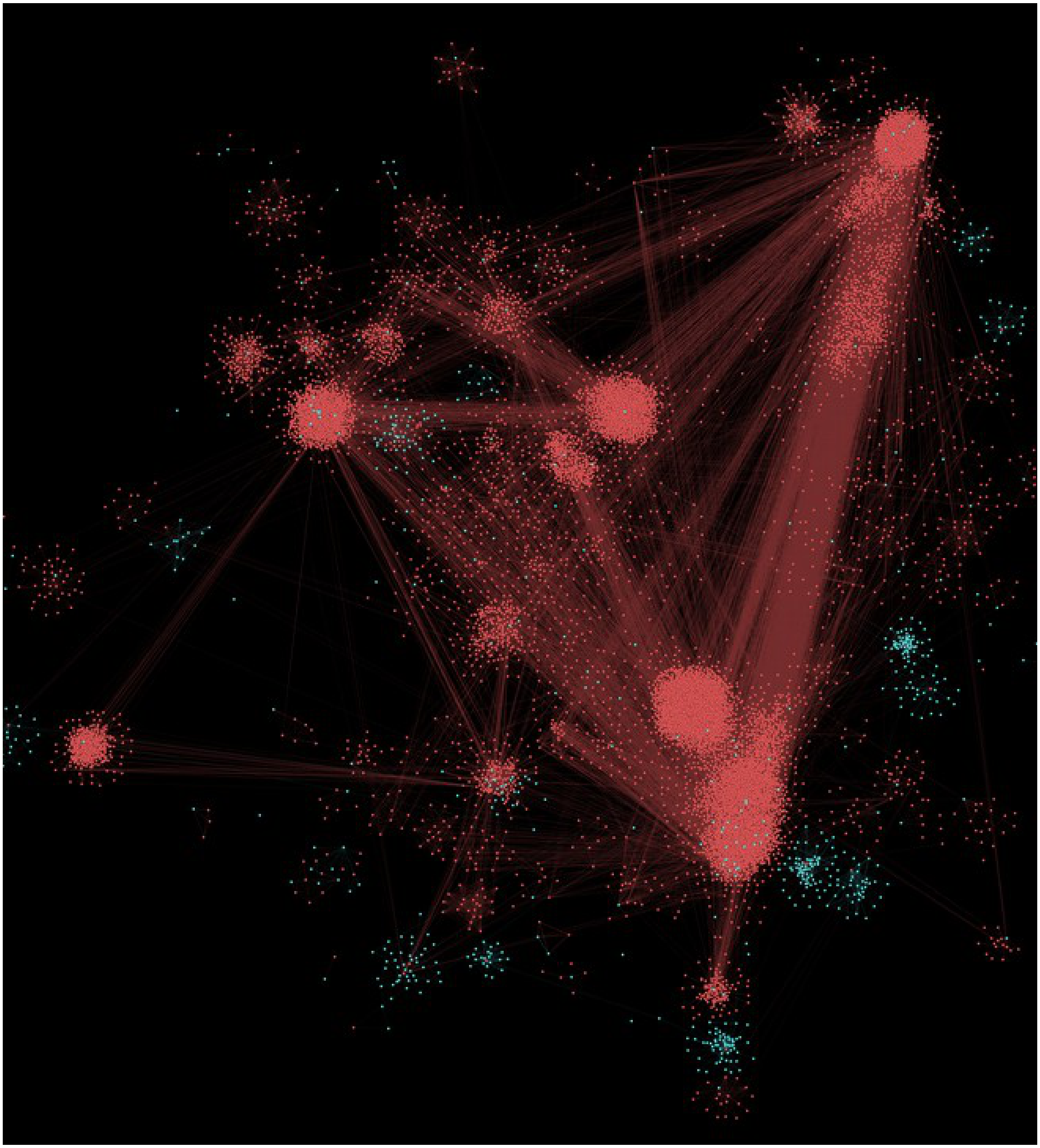


Figure 19. Gephi Subject to Relation network graph. A network graph of Subject to Relation, using the OpenOrd layout and coloured by type.

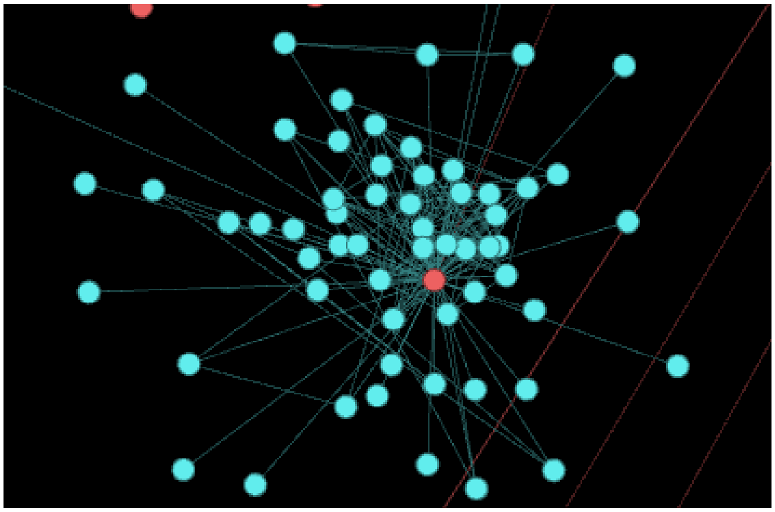


Figure 20. Cluster of many Relations around a single Subject term.

Clusters of blue nodes surrounding a single red node are instances where a collection uses only one subject term (see Figure 20). For example, the subject “faculty of Engineering” has many different isolated relations connected to it. Each is a collection of papers or fonds of a professor that has been assigned only a single subject term. This information is revealed by zooming in on the cluster and reading the labels or examining the underlying data.

A large clusters of red nodes with a single blue node nested inside, represents a collection with many subject headings that are highly specific and not repeated elsewhere (see Figure 21). These subjects are essentially self-referential, describing items within the context of the collection, but not connecting to more universal concepts.

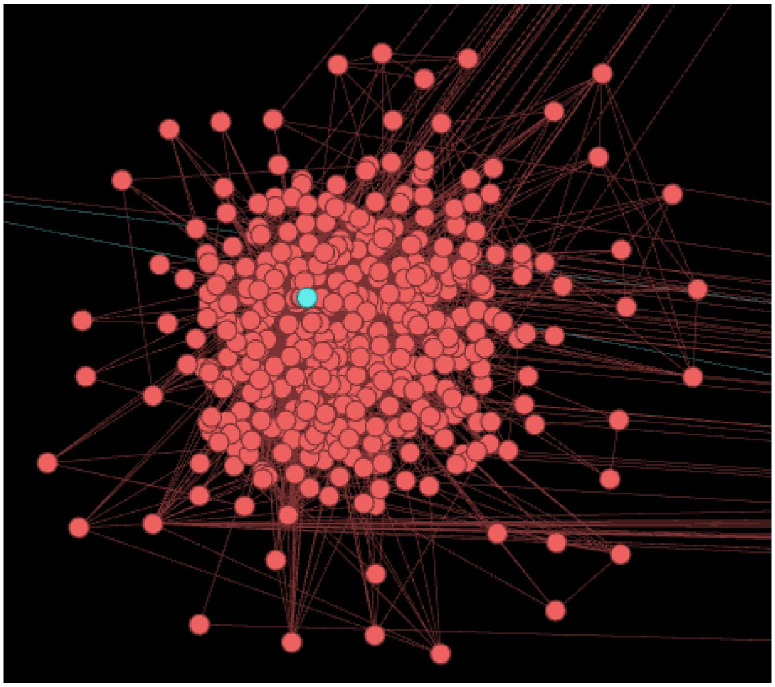


Figure 21. Cluster of many Subject terms around a single Relation.

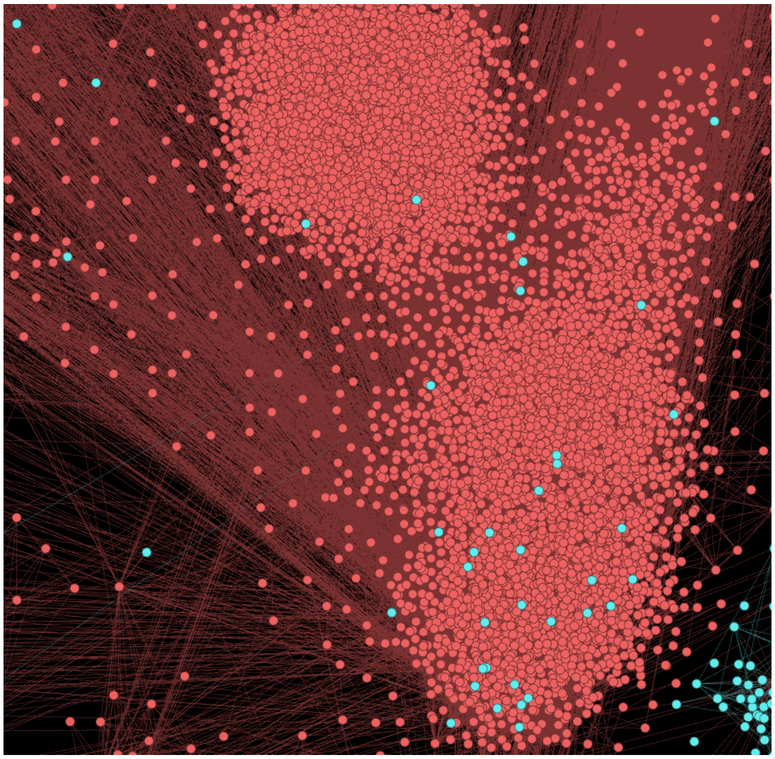


Figure 22. Cluster of evenly distributed nodes.

In contrast, the large clusters of red nodes with many blue nodes embedded are areas where more universal subject terms are applied that repeat across many different collections (see Figure 22). These differences point to the issue of balancing the use of contextualized versus standardized description. Locally contextualized metadata may best represent and provide findability within a single collection, fully representing its unique features. Standardized metadata conforms to a larger scope that gives up some individual features to insure a broader application and interoperability across the institution and beyond. Inspecting the underlying data from the network graph above, we can infer that in UBC Digital Collections the Chung Collection and the UBC Archives fonds deploy the most standardized use of subject terms.

This type of visualization reveals interesting structure with in the collections that is not comprehensible in a table or item level data. For managers of digital collections, it gives a sense of how subjects are being applied to better evaluate metadata practice across the repository. If we want to support browsing and serendipitous discovery across collections, isolated pockets of items can be discovered and evaluated. Managers can get a holistic sense of how specific subject terms are being applied and what relationships between items they are implying.

The process of creating these graphs takes time and the use is not as immediately practical as the survey plots. Their meaningfulness increases with the quality of the metadata, but it is also easy to produce graphs of relationships that have little significance. However, the elegant ease of navigation around Gephi network graphs haunts me—it is a truly beautiful and fascinating way to explore the metadata. How often would you spend happy hours staring at textual subject headings or scrolling through tables? This points to another use of these visualizations for users and managers to navigate and browse the repository. The network graph in this context provides a method of discovery for individual objects as well as revealing a sense of the coverage and character of the collections.

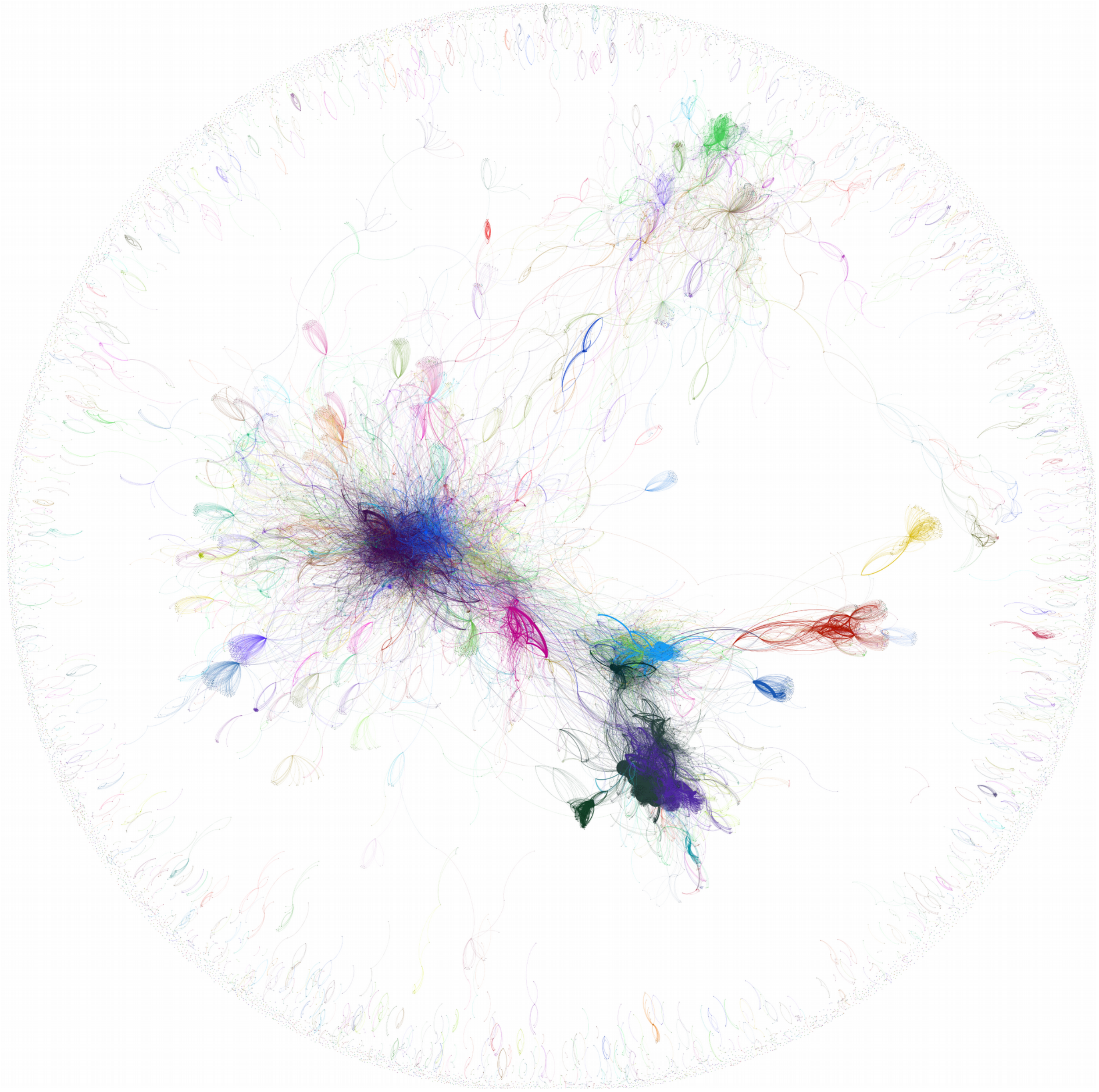


Figure 23. Gephi DC Subject network graph. Exported image for network graph using YifanHu layout and coloured using Chinese Whispers clustering.

For example, Figure 23 is an image exported from Gephi representing connections between Subject terms in UBC Digital Collections using the YifanHu force directed layout and coloured by the Chinese Whispers clustering algorithm.

The structure is similar to the previous example with a periphery of items having few relationships and a central cluster of more standardized terms. Screen shots do not do the Gephi graphs justice, but Figures 24-26 are a few details from the YifanHu layout shown on the previous page (note that the details look quite different than the exported PNG).

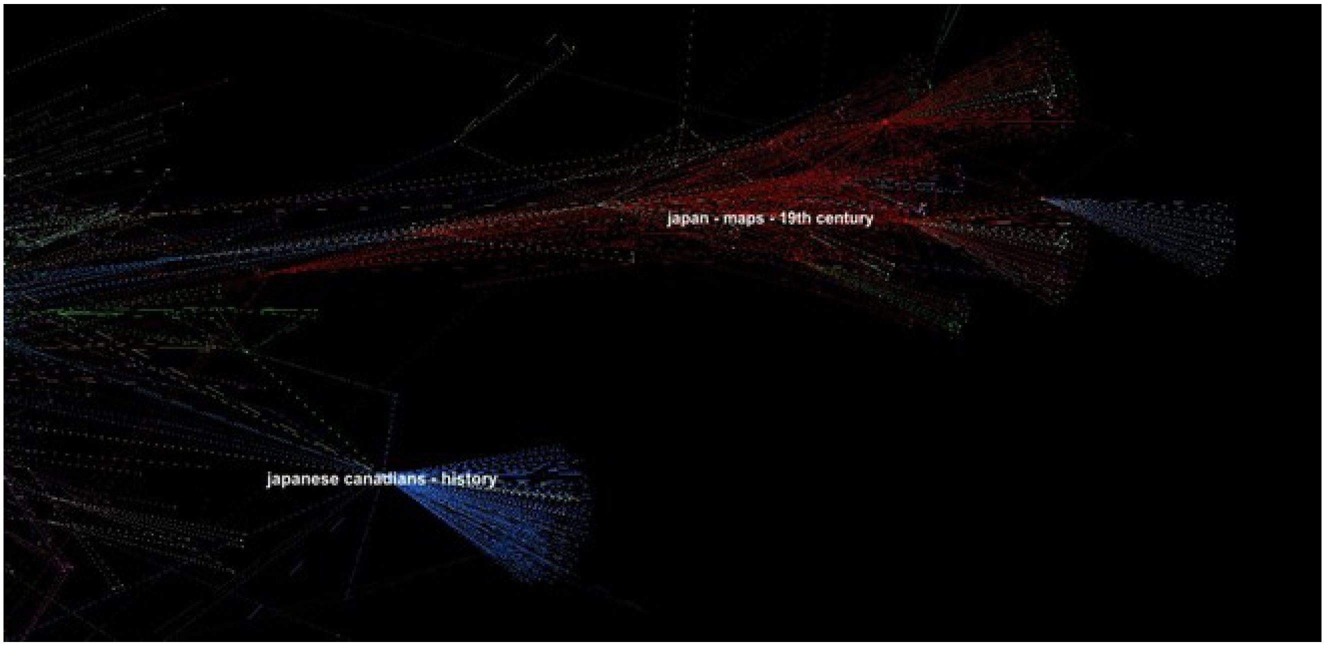


Figure 24. Detail of Gephi network graph.

The red arm extending out from the right clusters are subjects relating to Japan (see Figure 24). Most of the red nodes are subjects used to describe the Japanese Maps of the Tokugawa Era collection. However, terms used by number of other collections relating to Japan also get dragged in that direction such as the Japanese Canadian Photograph Collection.

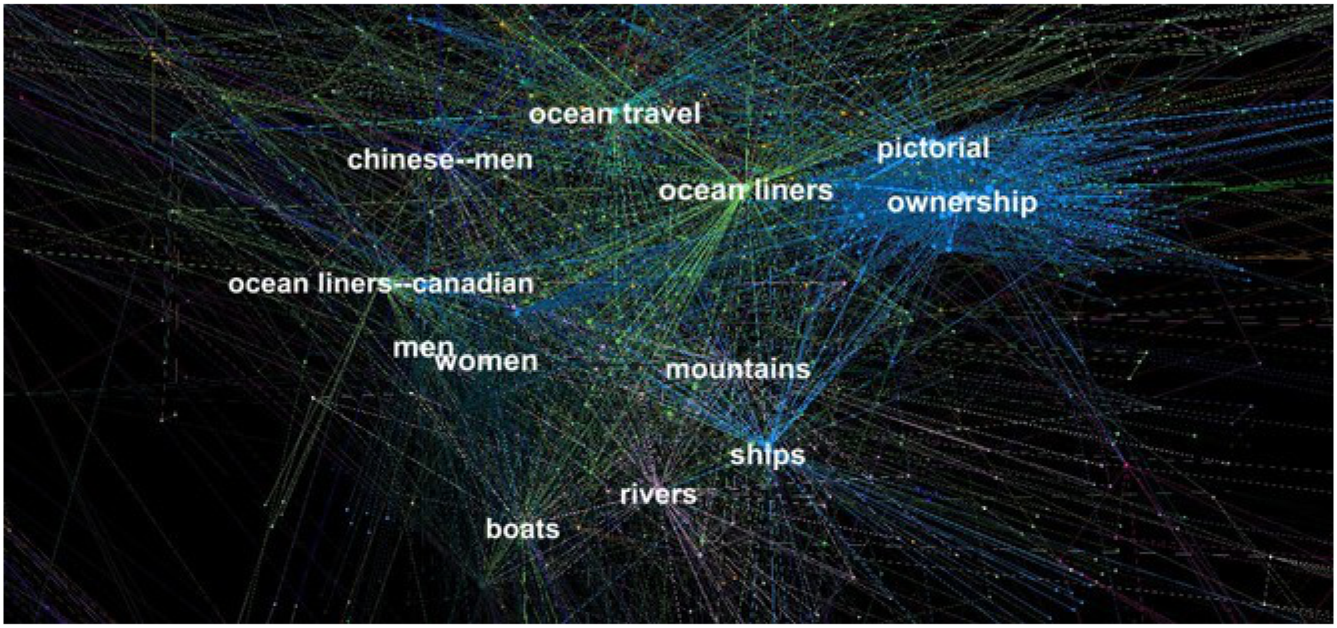


Figure 25. Detail of Gephi network graph.

The dense light blue cluster to the centre-right consists of universal subjects, such as man, woman, and tree (see Figure 25). These are most associated with the Chung Collections.

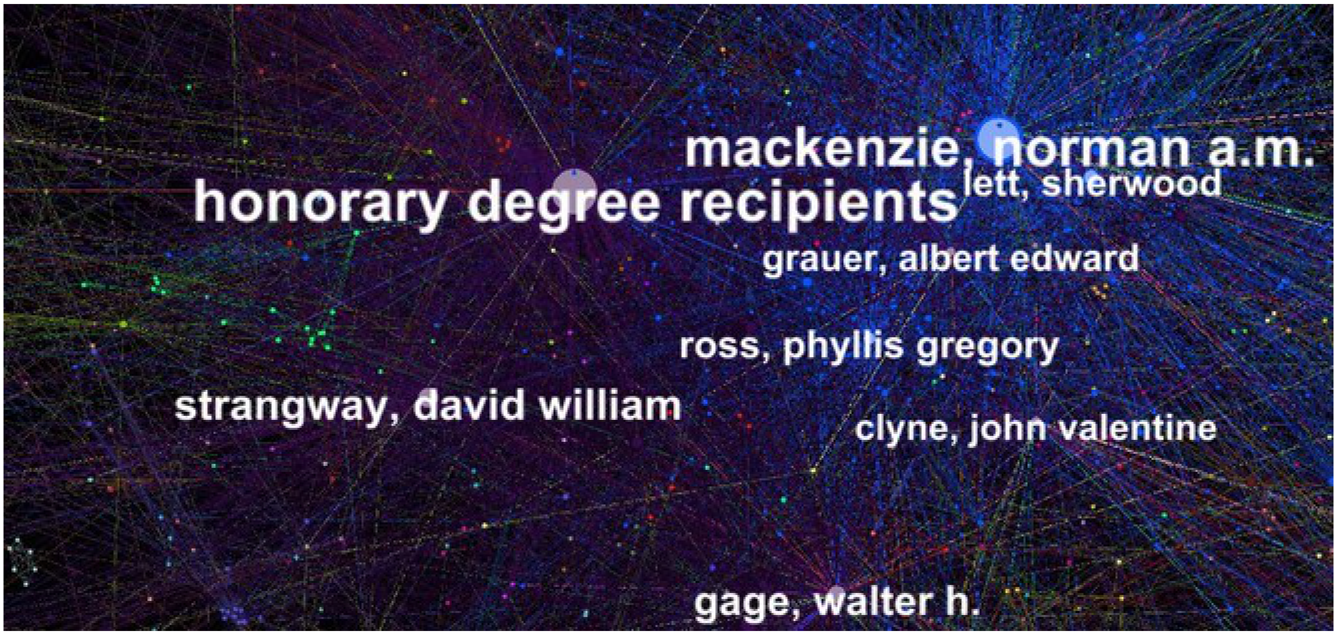


Figure 26. Detail of Gephi network graph.

The largest cluster, blue and purple to the left of centre, is made up mostly of names from items in the University Archives (see Figure 26).

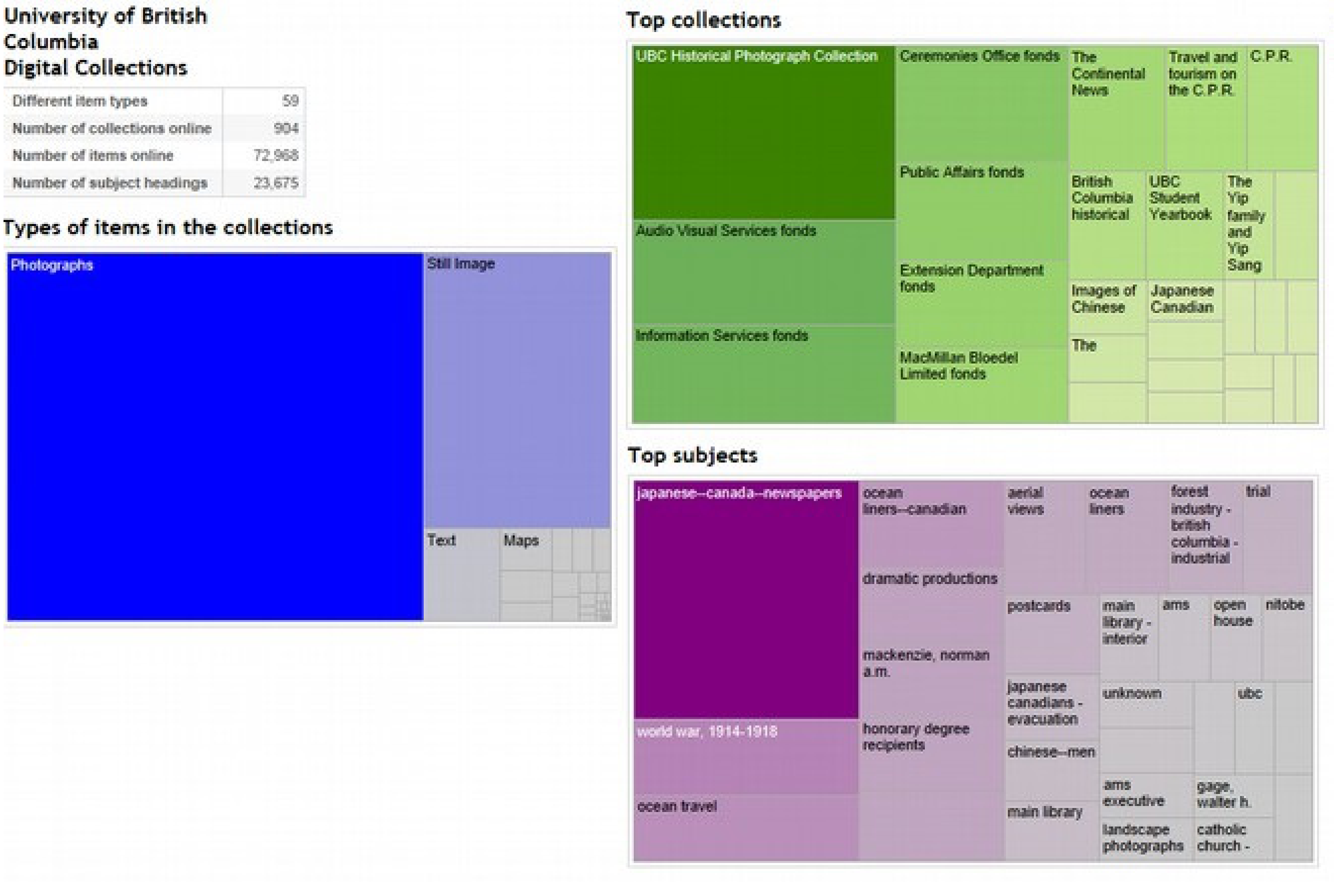
Extending these network graphs with a custom built tool could create a very interesting and useful representation of the overall collections. The visualization could provide insight into the repository's holdings, and act as a tool for faceted browsing. However, although the network graphs provide a powerful window into the relationships embedded in the repository, the quality of the representation is always limited by the consistency of the underlying metadata.

5. Dashboard

The final aim of this project was to create a simple representation that can orient website visitors to the repository. Since users often arrive from web search with no introduction to the institution, a quick glance should give a sense of context, scope, strengths, and character of the collections. Previous articles have discussed the dashboard concept for museums, such as one created for the Indianapolis Museum of Art (Urban, Twidale, & Adamczyk, 2010). This dashboard focuses on usage and funding statistics and was represented on a single page (http://dashboard.imamuseum.org). In this form, the dashboard is a valuable tool of institutional transparency, but not very descriptive of the collections themselves.

In contrast, I imagine a small icon that could appear in the corner of the uniform headers as miniaturized "map" of the collections. The icon would provide basic statistics about the repository. Clicking on any of the categories would bring up a visualization of the collections faceted on the field. While implementing a full version of this concept is beyond the scope of this project, I built a basic demonstration using Tableau to get a sense of what could be done. I imported the refined harvested metadata of UBC Digital Collections into Tableau. I created calculated fields counting the distinct values in dc_identifier, dc_subject, dc_relation, and dc_type and put these values in a simple table view.

Next, I created tree maps for dc_relation, dc_type, and dc_subject by number of items. To make the tree maps more readable, I filtered each to include only values with large counts, i.e. the "Top" entries. I then dragged the worksheets onto a dashboard (see Figure 27). Thanks to Tableau's efficient interface, this process took only about ten minutes.



University of British Columbia Digital Collections

Different item types	59
Number of collections online	904
Number of items online	72,968
Number of subject headings	23,675

Figure 27. Test dashboard on Tableau.

This simple visualization gives the disembodied digital item more sense of context and relationship to the institution. A slickly formatted web implementation would be beautiful, orient the user, and provide a new way of navigating the structures of the repository. As digital collections get larger and users complete more basic research online, better browsing capabilities help users go beyond the keyword search.

However, again as a word of caution, these representations are only as accurate as the quality of the metadata—which returns us to the original aim to efficiently evaluate and improve existing metadata for the repository as a whole. As this project demonstrates, visual analysis has the potential to open up metadata from a new, more holistic perspective—freeing us from the item-level tabular view to grasp the larger patterns and structures in the data. Seeing metadata will be a valuable tool for management, analysis, and exploration.

References:

Dushay, N., & Hillmann, D. I. (2003). Analyzing Metadata for Effective Use and Re-Use. Paper presented at DC-2003 Conference, Seattle, Wa, Oct 2003. Retrieved from http://hdl.handle.net/1813/7896.

Fenlon, K., Efron, M., & Organisciak, P. (2012). Tooling the Aggregator's Workbench: Metadata Visualization through Statistical Text Analysis. *Proceedings of the American Society for Information Science and Technology*, 49 (1), 1-10.

Levallois, C. (n.d.). Excel/csv converter to network. Retrieved from https://marketplace.gephi.org/plugin/excel-csv-converter-to-network.

Nichols, D. M., Paynter, G. W., Chan, C.H., Bainbridge, D., McKay, D., Twidale, M. B., & Blandford, A. (2009). Experiences in Deploying Metadata Analysis Tools for Institutional Repositories. *Cataloging & Classification Quarterly*, 47 (3-4), 229-248. DOI: 10.1080/01639370902737281.

Nichols, D. M., Chan, C.H., Bainbridge, D., McKay, D., & Twidale, M. B. (2008). A Lightweight Metadata Quality Tool. *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*, 385-388. DOI: 10.1145/1378889.1378957

Open Archives Initiative. (2002). The Open Archives Initiative Protocol for Metadata Harvesting. Retrieved from http://www.openarchives.org/OAI/openarchivesprotocol.html.

OpenRefine Wiki (n.d.). Clustering In Depth. Retrieved from https://github.com/OpenRefine/OpenRefine/wiki/Clustering-In-Depth.

Shreeves, S. L., Knutson, E. M., Stvilia, B., Palmer, C. L., Twidale, M. B., & Cole,

See Also:
2015 - Spring

T. W. (2005). Is 'Quality' Metadata 'Shareable' Metadata? The Implications of Local Metadata Practices for Federated Collections. *Proceedings of the 12th National Conference of ACRL*, 223-237. Retrieved from https://www.ideals.illinois.edu/bitstream/handle/2142/145/shreeves05.pdf.

Urban, R. J., Twidale, M. B., and Adamczyk, P. (2010). Designing and Developing a Collections Dashboard. *Proceedings of Museums and the Web 2010*. Retrieved from http://www.archimuse.com/mw2010/papers/urban/urban.html.

UBC Library. (n.d.). UBC Library Digital Collections. Retrieved from http://digitalcollections.library.ubc.ca.

Ware, C. (2013). *Information Visualization: Perception for Design*. New York: Morgan Kaufmann.

Westbrook, R. N., Johnson, D., Carter, K., & Lockwood, A. (2012). Metadata Clean Sweep: A Digital Library Audit Project. *D-Lib*, 18 (5/6). Retrieved from http://www.dlib.org/dlib/may12/westbrook/05westbrook.html.